Studies in Computational Intelligence 605

Stan Matwin
Jan Mielniczuk *Editors*

Challenges in Computational Statistics and Data Mining

Springer

Studies in Computational Intelligence

Volume 605

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series “Studies in Computational Intelligence” (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/7092

Stan Matwin · Jan Mielniczuk
Editors

Challenges in Computational Statistics and Data Mining

Springer

Editors

Stan Matwin
Faculty of Computer Science
Dalhousie University
Halifax, NS
Canada

Jan Mielniczuk
Institute of Computer Science
Polish Academy of Sciences
Warsaw
Poland

and

Warsaw University of Technology
Warsaw
Poland

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-319-18780-8 ISBN 978-3-319-18781-5 (eBook)
DOI 10.1007/978-3-319-18781-5

Library of Congress Control Number: 2015940970

Springer Cham Heidelberg New York Dordrecht London

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Preface

This volume contains 19 research papers belonging, roughly speaking, to the areas of computational statistics, data mining, and their applications. Those papers, all written specifically for this volume, are their authors' contributions to honour and celebrate Professor Jacek Koronacki on the occcasion of his 70th birthday. The volume is the brain-child of Janusz Kacprzyk, who has managed to convey his enthusiasm for the idea of producing this book to us, its editors. Books related and often interconnected topics, represent in a way Jacek Koronacki's research interests and their evolution. They also clearly indicate how close the areas of computational statistics and data mining are.

Mohammad Reza Bonyadi and Zbigniew Michalewicz in their article "Evolutionary Computation for Real-world Problems" describe their experience in applying Evolutionary Algorithms tools to real-life optimization problems. In particular, they discuss the issues of the so-called multi-component problems, the investigation of the feasible and the infeasible parts of the search space, and the search bottlenecks.

Susanne Bornelöv and Jan Komorowski "Selection of Significant Features Using Monte Carlo Feature Selection" address the issue of significant features detection in Monte Carlo Feature Selection method. They propose an alternative way of identifying relevant features based on approximation of permutation p-values by normal p-values and they compare its performance with the performance of built-in selection method.

In his contribution, Łukasz Dębowski "Estimation of Entropy from Subword Complexity" explores possibilities of estimating block entropy of stationary ergodic process by means of word complexity i.e. approximating function $f(k|w)$ which for a given string w yields the number of distinct substrings of length k. He constructs two estimates and shows that the first one works well only for iid processes with uniform marginals and the second one is applicable for much broader class of so-called properly skewed processes. The second estimator is used to corroborate Hilberg's hypothesis for block length no larger than 10.

Maik Döring, László Györfi and Harro Walk "Exact Rate of Convergence of Kernel-Based Classification Rule" study a problem in nonparametric classification concerning excess error probability for kernel classifier and introduce its decomposition into estimation error and approximation error. The general formula is provided for the approximation and, under a weak margin condition, its tight version.

Michał Dramiński in his exposition "ADX Algorithm for Supervised Classification" discusses a final version of rule-based classifier ADX. It summarizes several years of the author's research. It is shown in experiments that inductive methods may work better or on par with popular classifiers such as Random Forests or Support Vector Machines.

Olgierd Hryniewicz "Process Inspection by Attributes Using Predicted Data" studies an interesting model of quality control when instead of observing quality of inspected items directly one predicts it using values of predictors which are easily measured. Popular data mining tools such as linear classifiers and decision trees are employed in this context to decide whether and when to stop the production process.

Szymon Jaroszewicz and Łukasz Zaniewicz "Székely Regularization for Uplift Modeling" study a variant of uplift modeling method which is an approach to assess the causal effect of an applied treatment. The considered modification consists in incorporating Székely regularization into SVM criterion function with the aim to reduce bias introduced by biased treatment assignment. They demonstrate experimentally that indeed such regularization decreases the bias.

Janusz Kacprzyk and Sławomir Zadrożny devote their paper "Compound Bipolar Queries: A Step Towards an Enhanced Human Consistency and Human Friendliness" to the problem of querying of databases in natural language. The authors propose to handle the inherent imprecision of natural language using a specific fuzzy set approach, known as compound bipolar queries, to express imprecise linguistic quantifiers. Such queries combine negative and positive information, representing required and desired conditions of the query.

Miłosz Kadziński, Roman Słowiński, and Marcin Szeląg in their paper "Dominance-Based Rough Set Approach to Multiple Criteria Ranking with Sorting-Specific Preference Information" present an algorithm that learns ranking of a set of instances from a set of pairs that represent user's preferences of one instance over another. Unlike most learning-to-rank algorithms, the proposed approach is highly interactive, and the user has the opportunity to observe the effect of their preferences on the final ranking. The algorithm is extended to become a multiple criteria decision aiding method which incorporates the ordinal intensity of preference, using a rough-set approach.

Marek Kimmel "On Things Not Seen" argues in his contribution that frequently in biological modeling some statistical observations are indicative of phenomena which logically should exist but for which the evidence is thought missing. The claim is supported by insightful discussion of three examples concerning evolution, genetics, and cancer.

Mieczysław Kłopotek, Sławomir Wierzchoń, Robert Kłopotek and Elżbieta Kłopotek in "Network Capacity Bound for Personalized Bipartite PageRank" start from a simplification of a theorem for personalized random walk in an unimodal graph which is fundamental to clustering of its nodes. Then they introduce a novel notion of Bipartite PageRank and generalize the theorem for unimodal graphs to this setting.

Marzena Kryszkiewicz devotes her article "Dependence Factor as a Rule Evaluation Measure" to the presentation and discussion of a new evaluation measure for evaluation of associations rules. In particular, she shows how the dependence factor realizes the requirements for interestingness measures postulated by Piatetsky-Shapiro, and how it addresses some of the shortcomings of the classical certainty factor measure.

Adam Krzyżak "Recent Results on Nonparametric Quantile Estimation in a Simulation Model" considers a problem of quantile estimation of the random variable $m(X)$ where X has a given density by means of importance sampling using a regression estimate of m. It is shown that such yields a quantile estimator with a better asymptotic properties than the classical one. Similar results are valid when recursive Robbins-Monro importance sampling is employed.

The contribution of Błażej Miasojedov, Wojciech Niemiro, Jan Palczewski, and Wojciech Rejchel in "Adaptive Monte Carlo Maximum Likelihood" deal with approximation to the maximum likelihood estimator in models with intractable constants by adaptive Monte Carlo method. Adaptive importance sampling and a new algorithm which uses resampling and MCMC is investigated. Among others, asymptotic results, such that consistency and asymptotic law of the approximative ML estimators of the parameter are proved.

Jan Mielniczuk and Paweł Teisseyre in "What do We Choose When We Err? Model Selection and Testing for Misspecified Logistic Regression Revisited" consider common modeling situation of fitting logistic model when the actual response function is different from logistic one and provide conditions under which Generalized Information Criterion is consistent for set t^* of the predictors pertaining to the Kullback-Leibler projection of true model t. The interplay between t and t^* is also discussed.

Mirosław Pawlak in his contribution "Semiparametric Inference in Identification of Block-Oriented Systems" gives a broad overview of semiparametric statistical methods used for identification in a subclass of nonlinear-dynamic systems called block oriented systems. They are jointly parametrized by finite-dimensional parameters and an infinite-dimensional set of nonlinear functional characteristics. He shows that using semiparametric approach classical nonparametric estimates are amenable to the incorporation of constraints and avoid high-dimensionality/high-complexity problems.

Marina Sokolova and Stan Matwin in their article "Personal Privacy Protection in Time of Big Data" look at some aspects of data privacy in the context of big data analytics. They categorize different sources of personal health information and emphasize the potential of Big Data techniques for linking of these various sources. Among others, the authors discuss the timely topic of inadvertent disclosure of personal health information by people participating in social networks discussions.

Jerzy Stefanowski in his article "Dealing with Data Difficulty Factors while Learning from Imbalanced Data" provides a thorough review of the approaches to learning classifiers in the situation when one of the classes is severely underrepresented, resulting in a skewed, or imbalanced distribution. The article presents all the existing methods and discusses their advantages and shortcomings, and recommends their applicability depending on the specific characteristics of the imbalanced learning task.

In his article James Thompson "Data Based Modeling" builds a strong case for a data-based modeling using two examples: one concerning portfolio management and second being the analysis of hugely inadequate action of American health service to stop AIDS epidemic. The main tool in the analysis of the first example is an algorithm called MaxMedian Rule developed by the author and L. Baggett.

We are very happy that we were able to collect in this volume so many contributions intimately intertwined with Jacek's research and his scientific interests. Indeed, he is one of the authors of Monte Carlo Feature Selection system which is discussed here and widely contributed to nonparametric curve estimation and classification (subject of Döring et al. and Krzyżak's paper). He started his career with research in optimization and stochastic approximation—the themes being addressed in Bonyadi and Michalewicz as well as in Miasojedow et al. papers. He held long-lasting interests in Statistical Process Control discussed by Hryniewicz. He also has, as the contributors to this volume and his colleagues from Rice University, Thompson and Kimmel, keen interests in methodology of science and stochastic modeling.

Jacek Koronacki has been not only very active in research but also has generously contributed his time to the Polish and international research communities. He has been active in the International Organization of Standardization and in the European Regional Committee of the Bernoulli Society. He has been and is a longtime director of Institute of Computer Science of Polish Academy of Sciences in Warsaw. Administrative work has not prevented him from being an active researcher, which he continues up to now. He holds unabated interests in new developments of computational statistics and data mining (one of the editors vividly recalls learning about Székely distance, also appearing in one of the contributed papers here, from him). He has co-authored (with Jan Ćwik) the first Polish textbook in statistical Machine Learning. He exerts profound influence on the Polish data mining community by his research, teaching, sharing of his knowledge, refereeing, editorial work, and by exercising his very high professional standards. His friendliness and sense of humour are appreciated by all his colleagues and collaborators. In recognition of all his achievements and contributions, we join the authors of all the articles in this volume in dedicating to him this book as an expression of our gratitude. Thank you, Jacku; dziękujemy.

We would like to thank all the authors who contributed to this endeavor, and the Springer editorial team for perfect editing of the volume.

Ottawa, Warsaw, March 2015

Stan Matwin
Jan Mielniczuk

Contents

Evolutionary Computation for Real-World Problems

Mohammad Reza Bonyadi and Zbigniew Michalewicz

Abstract In this paper we discuss three topics that are present in the area of real-world optimization, but are often neglected in academic research in evolutionary computation community. First, problems that are a combination of several interacting sub-problems (so-called multi-component problems) are common in many real-world applications and they deserve better attention of research community. Second, research on optimisation algorithms that focus the search on the edges of feasible regions of the search space is important as high quality solutions usually are the boundary points between feasible and infeasible parts of the search space in many real-world problems. Third, finding bottlenecks and best possible investment in real-world processes are important topics that are also of interest in real-world optimization. In this chapter we discuss application opportunities for evolutionary computation methods in these three areas.

1 Introduction

The Evolutionary Computation (EC) community over the last 30 years has spent a lot of effort to design optimization methods (specifically Evolutionary Algorithms, EAs) that are well-suited for hard problems—problems where other methods usually

M.R. Bonyadi (✉) · Z. Michalewicz
Optimisation and Logistics, The University of Adelaide, Adelaide, Australia
e-mail: mrbonyadi@cs.adelaide.edu.au

Z. Michalewicz
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
e-mail: zbyszek@cs.adelaide.edu.au
url: http://cs.adelaide.edu.au/~optlog/

Z. Michalewicz
Polish-Japanese Institute of Information Technology, Warsaw, Poland

Z. Michalewicz
Chief of Science, Complexica, Adelaide, Australia

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_1

fail [36]. As most real-world problems[1] are very hard and complex, with nonlinearities and discontinuities, complex constraints and business rules, possibly conflicting objectives, noise and uncertainty, it seems there is a great opportunity for EAs to be used in this area.

Some researchers investigated features of real-world problems that served as reasons for difficulties of EAs when applied to particular problems. For example, in [53] the authors identified several such reasons, including premature convergence, ruggedness, causality, deceptiveness, neutrality, epistasis, and robustness, that make optimization problems hard to solve. It seems that these reasons are either related to the landscape of the problem (such as ruggedness and deceptiveness) or the optimizer itself (like premature convergence and robustness) and they are not focusing on the nature of the problem. In [38], a few main reasons behind the hardness of real-world problems were discussed; that included: the size of the problem, presence of noise, multi-objectivity, and presence of constraints. Apart from these studies on features related to the real-world optimization, there have been EC conferences (e.g. GECCO, IEEE CEC, PPSN) during the past three decades that have had special sessions on "real-world applications". The aim of these sessions was to investigate the potentials of EC methods in solving real-world optimization problems.

Consequently, most of the features discussed in the previous paragraph have been captured in optimization benchmark problems (many of these benchmark problems can be found in OR-library[2]). As an example, the size of benchmark problems has been increased during the last decades and new benchmarks with larger problems have appeared: knapsack problems (KP) with 2,500 items or traveling salesman problems (TSP) with more than 10,000 cities, to name a few. Noisy environments have been already defined [3, 22, 43] in the field of optimization, in both continuous and combinatorial optimization domain (mainly from the operations research field), see [3] for a brief review on robust optimization. Noise has been considered for both constraints and objective functions of optimization problems and some studies have been conducted on the performance of evolutionary optimization algorithms with existence of noise; for example, stochastic TSP or stochastic vehicle routing problem (VRP). We refer the reader to [22] for performance evaluation of evolutionary algorithms when the objective function is noisy. Recently, some challenges to deal with continuous space optimization problems with noisy constraints were discussed and some benchmarks were designed [43]. Presence of constraints has been also captured in benchmark problems where one can generate different problems with different constraints, for example Constrained VRP, (CVRP). Thus, the expectation is, after capturing all of these pitfalls and addressing them (at least some of them), EC optimization methods should be effective in solving real-world problems.

However, after over 30 years of research, tens of thousands of papers written on Evolutionary Algorithms, dedicated conferences (e.g. GECCO, IEEE CEC, PPSN),

[1] By real-world problems we mean problems which are found in some business/industry on daily (regular) basis. See [36] for a discussion on different interpretations of the term "real-world problems".

[2] Available at: http://people.brunel.ac.uk/~mastjjb/jeb/info.html.

dedicated journals (e.g. Evolutionary Computation Journal, IEEE Transactions on Evolutionary Computation), special sessions and special tracks on most AI-related conferences, special sessions on real-world applications, etc., still it is not that easy to find EC-based applications in real-world, especially in real-world supply chain industries.

There are several reasons for this mismatch between the efforts of hundreds of researchers who have been making substantial contribution to the field of Evolutionary Computation over many years and the number of real-world applications which are based on concepts of Evolutionary Algorithms—these are discussed in detail in [37]. In this paper we summarize our recent efforts (over the last two years) to close the gap between research activities and practice; these efforts include three research directions:

- Studying multi-component problems [7]
- Investigating boundaries between feasible and infeasible parts of the search space [5]
- Examining bottlenecks [11].

The paper is based on our four earlier papers [5, 7, 9, 11] and is organized as follows. We start with presenting two real-world problems (Sect. 2) so the connection between presented research directions and real-world problems is apparent. Sections 3–5 summarize our current research on studying multi-component problems, investigating boundaries between feasible and infeasible parts of the search space, and examining bottlenecks, respectively. Section 6 concludes the paper.

2 Example Supply Chains

In this section we explain two real-world problems in the field of supply chain management. We refer to these two examples further in the paper.

Transportation of water tank The first example relates to optimization of the transportation of water tanks [21]. An Australian company produces water tanks with different sizes based on some *orders* coming from its customers. The number of customers per month is approximately 10,000; these customers are in different locations, called *stations*. Each customer orders a water tank with specific characteristics (including size) and expects to receive it within a period of time (usually within 1 month). These water tanks are carried to the stations for delivery by a fleet of trucks that is operated by the water tank company. These trucks have different characteristics and some of them are equipped with trailers. The company proceeds in the following way. A subset of orders is selected and assigned to a truck and the delivery is scheduled in a limited period of time. Because the tanks are empty and of different sizes they might be packed inside each other in order to maximize trucks load in a trip. A bundled tank must be unbundled at special sites, called *bases*, before the tank delivery to stations. Note that there might exist several bases close to the stations where the tanks are going to be delivered and selecting different bases affects the best overall achievable solution. When the tanks are unbundled at a base, only some of them fit in the truck as they require more space. The truck is loaded with a subset of these tanks and deliver them to their corresponding stations for delivery. The remaining tanks are kept in the base until the truck gets back and loads them again to continue the delivery process.

The aim of the optimizer is to divide all tanks ordered by customers into subsets that are bundled and loaded in trucks (possibly with trailers) for delivery. Also, the optimizer needs to determine an exact routing for bases and stations for unbundling and delivery activities. The objective is to maximize the profit of the delivery at the end of the time period. This total profit is proportional to the ratio between the total prices of delivered tanks to the total distance that the truck travels.

Each of the mentioned procedures in the tank delivery problem (subset selection, base selection, and delivery routing, and bundling) is just one component of the problem and finding a solution for each component in isolation does not lead us to the optimal solution of the whole problem. As an example, if the subset selection of the orders is solved optimally (the best subset of tanks is selected in a way that the price of the tanks for delivery is maximized), there is no guarantee that there exist a feasible bundling such that this subset fits in a truck. Also, by selecting tanks without considering the location of stations and bases, the best achievable solutions can still have a low quality, e.g. there might be a station that needs a very expensive tank but it is very far from the base, which actually makes delivery very costly. On the other hand, it is impossible to select the best routing for stations before selecting tanks without selection of tanks, the best solution (lowest possible tour distance) is to deliver nothing. Thus, solving each sub-problem in isolation does not necessarily lead us to the overall optimal solution.

Note also that in this particular case there are many additional considerations that must be taken into account for any successful application. These include scheduling of drivers (who often have different qualifications), fatigue factors and labor laws, traffic patterns on the roads, feasibility of trucks for particular segments of roads, and maintenance schedule of the trucks.

Mine to port operation The second example relates to optimizing supply-chain operations of a mining company: from mines to ports [31, 32]. Usually in mine to port operations, the mining company is supposed to satisfy customer orders to provide predefined amounts of products (the raw material is dig up in mines) by a particular due date (the product must be ready for loading in a particular port). A port contains a huge area, called *stockyard*, several places to berth the ships, called *berths*, and a waiting area for the ships. The stockyard contains some *stockpiles* that are single-product storage units with some capacity (mixing of products in stockpiles is not allowed). Ships arrive in ports (time of arrival is often approximate, due to weather conditions) to take specified products and transport them to the customers. The ships wait in the waiting area until the port manager assigns them to a particular berth. Ships apply a cost penalty, called *demurrage*, for each time unit while it is waiting to be berthed since its arrival. There are a few *ship loaders* that are assigned to each berthed ship to load it with demanded products. The ship loaders take products from appropriate stockpiles and load them to the ships. Note that, different ships have different product demands that can be found in more than one stockpile, so that scheduling different ship loaders and selecting different stockpiles result in different amount of time to fulfill the ships demand. The goal of the mine owner is to provide sufficient amounts of each product type to the stockyard. However, it is also in the interest of the mine owner to minimize costs associated with early (or late) delivery, where these are estimated with respect to the (scheduled) arrival of the ship. Because mines are usually far from ports, the mining company has a number of trains that are used to transport products from a mine to the port. To operate trains, there is a rail network that is (usually) rented by the mining company so that trains can travel between mines and ports. The owner of the rail network sets some constraints for the operation of trains for each mining company, e.g. the number of passing trains per day through each junction (called *clusters*) in the network is a constant (set by the rail network owner) for each mine company.

There is a number of *train dumpers* that are scheduled to unload the products from the trains (when they arrive at port) and put them in the stockpiles. The mine company schedules trains and loads them at mine sites with appropriate material and sends them to the port while respecting all constraints (the *train scheduling* procedure). Also, scheduling train dumpers to unload the trains and put the unloaded products in appropriate stockpiles (the *unload scheduling* procedure), scheduling the ships to berth (this called *berthing* procedure), and scheduling the ship loaders to take products from appropriate stockpiles and load the ships (the *loader scheduling* procedure) are the other tasks for the mine company. The aim is to schedule the ships and fill them with the required products (ship demands) so that the total demurrage applied by all ships is minimized in a given time horizon.

Again, each of the aforementioned procedures (train scheduling, unload scheduling, berthing, and loader scheduling) is one component of the problem. Of course each of these components is a hard problem to solve by its own. Apart from the complication in each component, solving each component in isolation does not lead us to an overall solution for the whole problem. As an example, scheduling trains to optimality (bringing as much product as possible from mine to port) might result in insufficient available capacity in the stockyard or even lack of adequate products for the ships that arrive unexpectedly early. That is to say, ship arrival times have uncertainty associated with them (e.g. due to seasonal variation in weather conditions), but costs are independent of this uncertainty. Also, the best plan for dumping products from trains and storing them in the stockyard might result in a low quality plan for the ship loaders and result in too much movement to load a ship.

Note that, in the real-world case, there were some other considerations in the problem such as seasonal factor (the factor of constriction of the coal), hatch plan of ships (each product should be loaded in different parts of the ship to keep the balance of the vessel), availability of the drivers of the ship loaders, switching times between changing the loading product, dynamic sized stockpiles, etc.

Both problems illustrate the main issues discussed in the remaining sections of this document, as (1) they consist of several inter-connected components, (2) their boundaries between feasible and infeasible areas of the search space deserve careful examination, and (3) in both problems, the concept of bottleneck is applicable.

3 Multi-component Problems

There are thousands of research papers addressing traveling salesman problems, job shop and other scheduling problems, transportation problems, inventory problems, stock cutting problems, packing problems, various logistic problems, to name but a few. While most of these problems are NP-hard and clearly deserve research efforts, it is not exactly what the real-world community needs. Let us explain.

Most companies run complex operations and they need solutions for problems of high complexity with several components (i.e. multi-component problems; recall examples presented in Sect. 2). In fact, real-world problems usually involve several smaller sub-problems (several components) that interact with each other and companies are after a solution for the whole problem that takes all components into account rather than only focusing on one of the components. For example, the issue of scheduling production lines (e.g. maximizing the efficiency or minimizing the cost) has direct relationships with inventory costs, stock-safety levels, replenishments strategies, transportation costs, delivery-in-full-on-time (DIFOT) to customers, etc., so it should not be considered in isolation. Moreover, optimizing one component of the operation may have negative impact on upstream and/or downstream activities. These days businesses usually need "global solutions" for their operations, not component solutions. This was recognized over 30 years ago by Operations Research (OR) community; in [1] there is a clear statement: *Problems require holistic treatment. They cannot be treated effectively by decomposing them analytically into separate problems to which optimal solutions are sought.* However, there are very few research efforts which aim in that direction mainly due to the lack of appropriate benchmarks or test cases availability. It is also much harder to work with a company on such global level as the delivery of successful software solution usually involves many other (apart from optimization) skills, from understanding the companys internal processes to complex software engineering issues.

Recently a new benchmark problem called the traveling thief problem (TTP) was introduced [7] as an attempt to provide an abstraction of multi-component problems with dependency among components. The main idea behind TTP was to combine two problems and generate a new problem which contains two components. The TSP and KP were combined because both of these problems were investigated for many years in the field of optimization (including mathematics, operations research, and computer science). TTP was defined as a thief who is going to steal m items from n cities and the distance of the cities ($d(i, j)$ the distance between cities i and j), the profit of each item (p_i), and the weight of the items (w_i) are given. The thief is carrying a limited-capacity knapsack (maximum capacity W) to collect the stolen items. The problem is asked for the best plan for the thief to visit all cities exactly once (traveling salesman problem, TSP) and pick the items (knapsack problem, KP) from these cities in a way that its total benefit is maximized. To make the two sub-problems dependent, it was assumed that the speed of the thief is affected by the current weight of the knapsack (W_c) so that the more item the thief picks, the slower he can run. A function $v : \mathbb{R} \rightarrow \mathbb{R}$ is given which maps the current weight of the knapsack to the speed of thief. Clearly, $v\,(0)$ is the maximum speed of the thief (empty knapsack) and $v\,(W)$ is the minimum speed of the thief (full knapsack). Also, it was assumed that the thief should pay some of the profit by the time he completes the tour (e.g. rent of the knapsack, r). The total amount that should be paid is a function of the tour time. The total profit of the thief is then calculated by

$$B = P - r \times T$$

where B is the total benefit, P is the aggregation of the profits of the picked items, and T is the total tour time.

Generating a solution for KP or TSP in TTP is possible without being aware of the current solution for the other component. In addition, each solution for TSP impacts the best quality that can be achieved in the KP component because of the impact on the pay back that is a function of travel time. Moreover, each solution for the KP component impacts the tour time for TSP as different items impact the speed of travel differently due to the variability of weights of items. Some test problems were generated for TTP and some simple heuristic methods have been also applied to the problem [44].

Note that for a given instance of TSP and KP different values of r and functions f result in different instances of TTPs that might be harder or easier to solve. As an example, for small values of r (relative to P), the value of $r \times T$ has a small contribution to the value of B. In an extreme case, when $r = 0$, the contribution of $r \times T$ is zero, which means that the best solution for a given TTP is equivalent to the best solution of the KP component, hence, there is no need to solve the TSP component at all. Also, by increasing the value of r (relative to P), the contribution of $r \times T$ becomes larger. In fact, if the value of r is very large then the impact of P on B becomes negligible, which means that the optimum solution of the TTP is very close to the optimum solution of the given TSP (see Fig. 1).

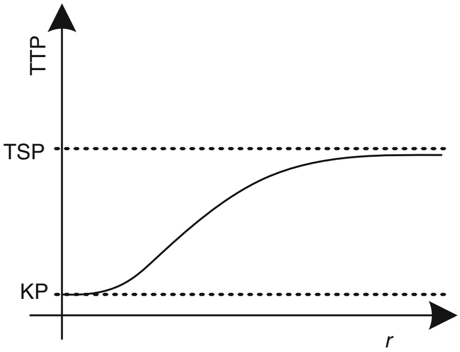


Fig. 1 Impact of the rent rate r on the TTP. For $r = 0$, the TTP solution is equivalent to the solution of KP, while for larger r the TTP solutions become closer to the solutions of TSP

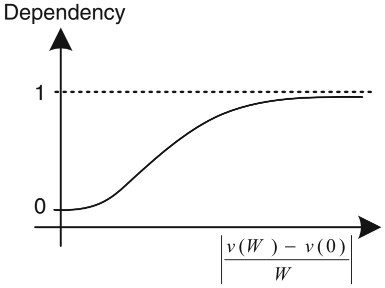


Fig. 2 How dependency between components is affected by speed (function v). When v does not drop significantly for different weights of picked items ($\left|\frac{v(W)-v(0)}{W}\right|$ is small), the two problems can be decomposed and solved separately. The value Dependency = 1 represents the two components are dependent while Dependency = 0 shows that two components are not dependent

The same analysis can be done for the function v. In fact, for a given TSP and KP different function v can result in different instances of TTPs that, as before, might be harder or easier. Let us assume that v is a decreasing function, i.e. picking items with positive weight causes drop or no change in the value of v. For a given list of items and cities, if picking an item does not affect the speed of the travel (i.e. $\left|\frac{v(W)-v(0)}{W}\right|$ is zero) significantly then the optimal solution of the TTP is the composition of the optimal solution of KP and TSP when they are solved separately. The reason is that, with this setting ($\left|\frac{v(W)-v(0)}{W}\right|$ is zero), picking more items does not change the time of the travel. As the value of $\left|\frac{v(W)-v(0)}{W}\right|$ grows, the TSP and KP become more dependent (picking items have more significant impact on the travel time); see Fig. 2.

As the value of $\left|\frac{v(W)-v(0)}{W}\right|$ grows, the speed of the travel drops more significantly by picking more items that in fact reduces the value of B significantly. In an extreme case, if $\left|\frac{v(W)-v(0)}{W}\right|$ is infinitely large then it would be better not to pick any item (the solution for KP is to pick no item) and only solve the TSP part as efficiently as possible. This has been also discussed in [10].

Recently, we generated some test instances for TTP and made them available [44] so that other researchers can also work along this path. The instance set contains 9,720 problems with different number of cities and items. The specification of the tour was taken from existing TSP problems in OR-Library. Also, we proposed three algorithms to solve those instances: one heuristic, one random search with local improvement, and one simple evolutionary algorithm. Results indicated that the evolutionary algorithm outperforms other methods to solve these instances. These test sets were also used in a competition in CEC2014 where participants were asked to come up with their algorithms to solve the instances. Two popular approaches emerged: combining different solvers for each sub-problem and creating one system for the overall problem.

Problems that require the combination of solvers for different sub-problems, one can find different approaches in the literature. First, in bi-level-optimization (and in the more general multi-level-optimization), one component is considered the dominant one (with a particular solver associated to it), and every now and then the other component(s) are solved to near-optimality or at least to the best extent possible by other solvers. In its relaxed form, let us call it "round-robin optimization", the optimization focus (read: CPU time) is passed around between the different solvers for the subcomponents. For example, this approach is taken in [27], where two heuristics are applied alternatingly to a supply-chain problem, where the components are (1) a dynamic lot sizing problem and (2) a pickup and delivery problem with time windows. However, in neither set-up did the optimization on the involved components commence in parallel by the solvers.

A possible approach to multi-component problems with presence of dependencies is based on the cooperative coevolution: a type of multi-population Evolutionary Algorithm [45]. Coevolution is a simultaneous evolution of several genetically isolated subpopulations of individuals that exist in a common ecosystem. Each subpopulation is called species and mate only within its species. In EC, coevolution can be of three types: competitive, cooperative, and symbiosis. In competitive coevolution, multiple species coevolve separately in such a way that fitness of individual from one species is assigned based on how good it competes against individuals from the other species. One of the early examples of competitive coevolution is the work by Hillis [20], where he applied a competitive predator-prey model to the evolution of sorting networks. Rosin and Belew [47] used the competitive model of coevolution to solve number of game learning problems including Tic-Tac-Toe, Nim and small version of Go. Cooperative coevolution uses divide and conquer strategy: all parts of the problem evolve separately; fitness of individual of particular species is assigned based on the degree of collaboration with individuals of other species. It seems that cooperative coevolution is a natural fit for multi-component problems with presence of dependencies. Individuals in each subpopulation may correspond to potential solutions for particular component, with its own evaluation function, whereas the global evaluation function would include dependencies between components. Symbiosis is another coevolutionary process that is based on living together of organisms of different species. Although this type appears to represent a more effective mechanism for automatic hierarchical models [19], it has not been studied in detail in the EC literature.

Additionally, feature-based analysis might be helpful to provide new insights and help in the design of better algorithms for multi-component problems. Analyzing statistical feature of classical combinatorial optimization problems and their relation to problem difficulty has gained an increasing attention in recent years [52]. Classical algorithms for the TSP and their success depending on features of the given input have been studied in [34, 41, 51] and similar analysis can be carried out for the knapsack problem. Furthermore, there are different problem classes of the knapsack problem which differ in their hardness for popular algorithms [33]. Understanding the features of the underlying sub-problems and how the features of interactions in a multi-component problem determine the success of different algorithms is an interesting topic for future research which would guide the development and selection of good algorithms for multi-component problems.

In the field of machine learning, the idea of using multiple algorithms to solve a problem in a better way has been used for decades. For example, ensemble methods—such as boosting, bagging, and stacking—use multiple learning algorithms to search the hypothesis space in different ways. In the end, the predictive performance of the combined hypotheses is typically better than the performances achieved by the constituent approaches.

Interestingly, transferring this idea into the optimization domain is not straightforward. While we have a large number of optimizers at our disposal, they are typically not general-purpose optimizers, but very specific and highly optimized for a particular class of problems, e.g., for the knapsack problem or the travelling salesperson problem.

4 Boundaries Between Feasible and Infeasible Parts of the Search Space

A constrained optimization problem (COP) is formulated as follows:

$$\text{find } x \in \mathcal{F} \subseteq \text{S} \subseteq R^D \text{ such that } \begin{cases} f(x) \leq f(y) & \text{for all } y \in \mathcal{F} \quad \text{(a)} \\ g_i(\text{x}) \leq 0 & \text{for } i = 1 \text{ to } q \quad \text{(b)} \\ h_i(x) = 0 & \text{for } i = q+1 \text{ to } m \quad \text{(c)} \end{cases} \tag{1}$$

where f, g_i, and h_i are real-valued functions on the search space S, q is the number of inequalities, and $m - q$ is the number of equalities. The set of all feasible points which satisfy constraints (b) and (c) are denoted by $\mathcal{F}$ [39]. The equality constraints are usually replaced by $|h_i(x)| - \sigma \leq 0$ where σ is a small value (normally set to 10^{-4}) [6]. Thus, a COP is formulated as

$$\text{find } x \in \mathcal{F} \subseteq \text{S} \subseteq R^D \text{ such that } \begin{cases} f(x) \leq f(y) & \text{for all } y \in \mathcal{F} \quad \text{(a)} \\ g_i(\text{x}) \leq 0 & \text{for } i = 1 \text{ to } m \quad \text{(b)} \end{cases} \tag{2}$$

where $g_i(x) = |h_i(x)| - \sigma$ for all $i \in \{q+1, \ldots, m\}$. Hereafter, the term COP refers to this formulation.

The constraint $g_i(x)$ is called *active* at the point x if the value of $g_i(x)$ is zero. Also, if $g_i(x) < 0$ then $g_i(x)$ is called *inactive* at x. Obviously, if x is feasible and at least one of the constraints is active at x, then x is on the boundary of the feasible and infeasible areas of the search space.

In many real-world COPs it is highly probable that some constraints are active at optimum points [49], i.e. some optimum points are on the edge of feasibility. The reason is that constraints in real-world problems often represent some limitations of resources. Clearly, it is beneficial to make use of some resources as much as possible, which means constraints are active at quality solutions. Presence of active constraints at the optimum points causes difficulty for many optimization algorithms to locate optimal solution [50]. Thus, it might be beneficial if the algorithm is able to focus the search on the edge of feasibility for quality solutions.

So it is assumed that there exists at least one active constraint at the optimum solution of COPs. We proposed [5] a new function, called Subset Constraints Boundary Narrower (SCBN), that enabled the search methods to focus on the boundary of feasibility with an adjustable thickness rather than the whole search space. SCBN is actually a function (with a parameter ε for thickness) that, for a point x, its value is smaller than zero if and only if x is feasible and the value of *at least one* of the constraints in *a given subset* of all constraint of the COP at the point x is within a predefined boundary with a specific thickness. By using SCBN in any COP, the feasible area of the COP is limited to the boundary of feasible area defined by SCBN, so that the search algorithms can only focus on the boundary. Some other extensions of SCBN are proposed that are useful in different situations. SCBN and its extensions are used in a particle swarm optimization (PSO) algorithm with a simple constraint handling method to assess if they are performing properly in narrowing the search on the boundaries.

A COP can be rewritten by combining all inequality constraints to form only one inequality constraint. In fact, any COP can be formulated as follows:

$$\text{find } x \in \mathcal{F} \subseteq \mathrm{S} \subseteq R^D \text{ such that } \begin{cases} f(x) \leq f(y) \text{ for all } y \in \mathcal{F} & \text{(a)} \\ M(\mathrm{x}) \leq 0 & \text{(b)} \end{cases} \tag{3}$$

where $M(x)$ is a function that combines all constraints $g_i(x)$ into one function. The function $M(x)$ can be defined in many different ways. The surfaces that are defined by different instances of $M(x)$ might be different. The inequality 3(b) should capture the feasible area of the search space. However, by using problem specific knowledge, one can also define $M(x)$ in a way that the area that is captured by $M(x) \leq 0$ only refers to a sub-space of the whole feasible area where high quality solutions might be found. In this case, the search algorithm can focus only on the captured area which is smaller than the whole feasible area and make the search more effective. A frequently-used [29, 48] instance of $M(x)$ is a function $K(x)$

$$K(x) = \sum_{i=1}^{m} \max\{g_i(x), 0\} \tag{4}$$

Clearly, the value of $K(x)$ is non-negative. $K(x)$ is zero if and only if x is feasible. Also, if $K(x) > 0$, the value of $K(x)$ represents the maximum violation value (called the *constraint violation* value).

As in many real-world COPs, there is at least one active constraint near the global best solution of COPs [49], some researchers developed operators to enable search methods to focus the search on the edges of feasibility. GENOCOP (GEnetic algorithm for Numerical Optimization for Constrained Optimization) [35] was probably the first genetic algorithm variant that applied boundary search operators for dealing with COPs. Indeed, GENOCOP had three mutations and three crossovers operators and one of these mutation operators was a boundary mutation which could generate a random point on the boundary of the feasible area. Experiments showed that the presence of this operator caused significant improvement in GENOCOP for finding optimum for problems which their optimum solution is on the boundary of feasible and infeasible area [35].

A specific COP was investigated in [40] and a specific crossover operator, called *geometric crossover*, was proposed to deal with that COP. The COP was defined as follows:

$$\begin{aligned} f(x) &= \left| \frac{\sum_{i=1}^{D} cos^4(x_i) - 2\prod_{i=1}^{D} cos^2(x_i)}{\sqrt{\sum_{i=1}^{D} i x_i^2}} \right| \\ g_1(x) &= 0.75 - \prod_{i=1}^{D} x_i \le 0 \\ g_2(x) &= \sum_{i=1}^{D} x_i - 0.75D \ \le 0 \end{aligned} \tag{5}$$

where $0 \le x_i \le 10$ for all i. Earlier experiments [23] shown that the value of the first constraint ($g_1(x)$) is very close to zero at the best known feasible solution for this COP. The geometric crossover was designed as $x_{new,j} = \sqrt{x_{1,i}x_{2,j}}$, where $x_{i,j}$ is the value of the jth dimension of the ith parent, and $x_{new,j}$ is the value of the jth dimension of the new individual. By using this crossover, if $g_1(\mathbf{x}_1) = g_1(\mathbf{x}_2) = 0$, then $g_1(\mathbf{x}_{new}) = 0$ (the crossover is *closed* under $g_1(x)$). It was shown that an evolutionary algorithm that uses this crossover is much more effective than an evolutionary algorithm which uses other crossover operators in dealing with this COP. In addition, another crossover operator was also designed [40], called *sphere crossover*, that was closed under the constraint $g(x) = \sum_{i=1}^{D} x_i^2 - 1$. In the sphere crossover, the value of the new offspring was generated by $x_{new,j} = \sqrt{\alpha x_{1,j}^2 + (1-\alpha) x_{2,j}^2}$, where $x_{i,j}$ is the value of the jth dimension of the ith parent, and both parents $\mathbf{x}_1$ and $\mathbf{x}_2$ are on $g(x)$. This operator could be used if $g(x)$ is the constraint in a COP and it is active on the optimal solution.

In [50] several different crossover operators closed under $g(x) = \sum_{i=1}^{D} x_i^2 - 1$ were discussed. These crossovers operators included repair, sphere (explained above), *curve*, and *plane* operators. In the repair operator, each generated solution was normalized and then moved to the surface of $g(x)$. In this case, any crossover and mutation could be used to generate offspring; however, the resulting offspring is moved (repaired) to the surface of $g(x)$. The curve operator was designed in a way that it could generate points on the *geodesic curves*, curves with minimum length on a surface, on $g(x)$. The plane operator was based on the selection of a plane which contains both parents and crosses the surface of $g(x)$. Any point on this intersection is actually on the surface of the $g(x)$ as well. These operators were incorporated into several optimization methods such as GA and Evolutionary Strategy (ES) and the results of applying these methods to two COPs were compared.

A variant of evolutionary algorithm for optimization of a water distribution system was proposed [54]. The main argument was that the method should be able to make use of information on the edge between infeasible and feasible area to be effective in solving the water distribution system problem. The proposed approach was based on an adapting penalty factor in order to guide the search towards the boundary of the feasible search space. The penalty factor was changed according to the percentage of the feasibility of the individuals in the population in such a way that there are always some infeasible solutions in the population. In this case, crossover can make use of these infeasible and feasible individuals to generate solutions on the boundary of feasible region.

In [28] a boundary search operator was adopted from [35] and added to an ant colony optimization (ACO) method. The boundary search was based on the fact that the line segment that connects two points x and y, where one of these points are infeasible and the other one is feasible, crosses the boundary of feasibility. A binary search can be used to search along this line segment to find a point on the boundary of feasibility. Thus, any pair of points (x, y), where one of them is infeasible and the other is feasible, represents a point on the boundary of feasibility. These points were moved by an ACO during the run. Experiments showed that the algorithm is effective in locating optimal solutions that are on the boundary of feasibility.

In [5] we generalized the definition of edges of feasible and infeasible space by introducing thickness of the edges. We also introduced a formulation that, for any given COP, it could generate another COP that the feasible area of the latter corresponds to the edges of feasibility of the former COP. Assume that for a given COP, it is known that *at least one* of the constraints in the set $\{g_{i\in\Omega}(x)\}$ is active at the optimum solution and the remaining constraints are satisfied at x, where $\Omega \subseteq \{1, 2, \ldots, m\}$. We defined $H_{\Omega,\varepsilon}(x)$ as follows:

$$H_{\Omega,\varepsilon}(x) = \max\left\{\left|\max_{i\in\Omega}\{g_i(x)\} + \varepsilon\right| - \varepsilon,\ \max_{i\notin\Omega}\{g_i(x)\}\right\} \tag{6}$$

where ε is a positive value. Obviously, $H_{\Omega,\varepsilon}(x) \leq 0$ if and only if at least one of the constraints in the subset Ω is active and the others are satisfied. The reason is that, the component $\left|\max_{i\in\Omega}\{g_i(x)\} + \varepsilon\right| - \varepsilon$ is negative if x is feasible and at least one of $g_{i\in\Omega}(x)$ is active. Also, the component $\max_{i\notin\Omega}\{g_i(x)\}$ ensures that the rest of constraints are satisfied. Note that active constraints are considered to have a value between 0 and -2ε, i.e., the value of 2ε represents the thickness of the edges. This formulation can restrict the feasible search space to only the edges so that optimization algorithms are enforced to search the edges. Also, it enabled the user to provide a list of active constraints so that expert knowledge can help the optimizer to converge faster to better solutions.

Clearly methodologies that focuses the search on the edges of feasible area are beneficial for optimization in real-world. As an example, in the mining problem described in Sect. 2, it is very likely that using all of the trucks, trains, shiploaders, and train dumpers to the highest capacity is beneficial for increasing throughput. Thus, at least one of these constraints (resources) is active, which means that searching the edges of feasible areas of the search space very likely leads us to high quality solutions.

5 Bottlenecks

Usually real-world optimization problems contain constraints in their formulation. The definition of constraints in management sciences is anything that limits a system from achieving higher performance versus its goal [17]. In the previous section we provided general formulation of a COP. As discussed in the previous section, it is believed that the optimal solution of most real-world optimization problems is found on the edge of a feasible area of the search space of the problem [49]. This belief is not limited to computer science, but it is also found in operational research (linear programming, LP) [12] and management sciences (theory of constraints, TOC) [30, 46] articles. The reason behind this belief is that, in real-world optimization problems, constraints usually represent limitations of availability of resources. As it is usually beneficial to utilize the resources as much as possible to achieve a high-quality solution (in terms of the objective value, f), it is expected that the optimal solution is a point where a subset of these resources is used as much as possible, i.e., $g_i\left(x^*\right) = 0$ for some i and a particular high-quality x^* in the general formulation of COPs [5]. Thus, the best feasible point is usually located where the value of these constraints achieves their maximum values (0 in the general formulation). The constraints that are active at the optimum solution can be thought of as *bottlenecks* that constrain the achievement of a better objective value [13, 30].

Decision makers in industries usually use some tools, known as decision support systems (DSS) [24], as a guidance for their decisions in different areas of their systems. Probably the most important areas that decision makers need guidance from DSS are: (1) optimizing schedules of resources to gain more benefit (accomplished by an optimizer in DSS), (2) identifying bottlenecks (accomplished by analyzing constraints in DSS), and (3) determining the best ways for future investments to improve their profits (accomplished by an analysis for removing bottlenecks,[3] known as what-if analysis in DSS). Such support tools are more readily available than one

[3] The term removing a bottleneck refers to the investment in the resources related to that bottleneck to prevent those resources from constraining the problem solver to achieve better objective values.

might initially think: for example, the widespread desktop application Microsoft Excel provides these via an add-in.[4]

Identification of bottlenecks and the best way of investment is at least as valuable as the optimization in many real-world problems from an industrial point of view because [18]: *An hour lost at a bottleneck is an hour lost for the entire system. An hour saved at a non-bottleneck is a mirage.* Industries are not only after finding the best schedules of the resources in their systems (optimizing the objective function), but they are also after understanding the tradeoffs between various possible investments and potential benefits.

During the past 30 years, evolutionary computation methodologies have provided appropriate tools as optimizers for decision makers to optimize their schedules. However, the last two areas (identifying bottlenecks and removing them) that are needed in DSSs seem to have remained untouched by EC methodologies while it has been an active research area in management and operations research.

There have been some earlier studies on identifying and removing bottlenecks [14, 16, 25, 30]. These studies, however, have assumed only linear constraints and they have related bottlenecks only to one specific property of resources (usually the availability of resources). Further, they have not provided appropriate tools to guide decision makers in finding the best ways of investments in their system so that their profits are maximized by removing the bottlenecks. In our recent work [11], we investigated the most frequently used bottleneck removing analysis (so-called average shadow prices) and identified its limitations. We argued that the root of these limitations can be found in the interpretation of constraints and the definition of bottlenecks. We proposed a more comprehensive definition for bottlenecks that not only leads us to design a more comprehensive model for determining the best investment in the system, but also addresses all mentioned limitations. Because the new model was multi-objective and might lead to the formulation of non-linear objective functions/constraints, evolutionary algorithms have a good potential to be successful on this proposed model. In fact, by applying multi-objective evolutionary algorithms to the proposed model, the solutions found represent points that optimize the objective function and the way of investment with different budgets at the same time.

Let us start with providing some background information on linear programming, the concept of shadow price, and bottlenecks in general. A Linear Programming (LP) problem is a special case of COP, where $f(x)$ and $g_i(x)$ are linear functions:

$$\text{find } x \text{ such that } z = \max c^T x \text{ subject to } Ax \leq b^T \tag{7}$$

where A is a $m \times d$ dimensional matrix known as *coefficients matrix*, m is the number of constraints, d is the number of dimensions, c is a d-dimensional vector, b is a m-dimensional vector known as Right Hand Side (RHS), $x \in \mathbb{R}^d$, and $x \geq 0$.

[4] http://tinyurl.com/msexceldss, last accessed 29th March 2014.

The shadow price (SP) for the ith constraint of this problem is the value of z when b_i is increased by one unit. This in fact refers to the best achievable solution if the RHS of the ith constraint was larger, i.e., there were more available resources of the type i [26].

The concept of SP in Integer Linear Programming (ILP) is different from the one in LP [13]. The definition for ILP is similar to the definition of LP, except that $x \in \mathbb{Z}^d$. In ILP, the concept of Average Shadow Price (ASP) was introduced [25]. Let us define the *perturbation function* $z_i(w)$ as follows:

$$\text{find } x \text{ such that } z_i(w) = \max c^T x \text{ subject to } a_i x \le b_i + w \; a_k x \le b_k \; \forall k \ne i \quad (8)$$

where a_i is the ith row of the matrix A and $x \ge 0$. Then, the ASP for the ith constraint is defined by $ASP_i = \sup\limits_{w>0} \left\{ \frac{(z_i(w) - z_i(0))}{w} \right\}$. ASP_i represents that if adding one unit of the resource i costs p and $p < ASP_i$, then it is beneficial (the total profit is increased) to buy w units of this resource. This information is very valuable for the decision maker as it is helpful for removing bottlenecks. Although the value of ASP_i refers to "buying" new resources, it is possible to similarly define a selling shadow price [25].

Several extensions of this ASP definition exist. For example, a set of resources is considered in [15] rather than only one resource at a time. There, it was also shown that ASP can be used in mixed integer LP (MILP) problems.

Now, let us take a step back from the definition of ASP in the context of ILP, and let us see how it fits into a bigger picture of resources and bottlenecks. As we mentioned earlier, constraints usually model availability of resources and limit the optimizers to achieve the best possible solution which maximizes (minimizes) the objective function [26, 30, 46]. Although finding the best solution with the current resources is valuable for decision makers, it is also valuable to explore opportunities to improve solutions by adding more resources (e.g., purchasing new equipment) [25]. In fact, industries are seeking the most efficient way of investment (removing the bottlenecks) so that their profit is improved the most.

Let us assume that the decision maker has the option of providing some additional resource of type i at a price p. It is clearly valuable if the problem solver can determine if adding a unit of this resource can be beneficial in terms of improving the best achievable objective value. It is not necessarily the case that adding a new resource of the type i improves the best achievable objective value. As an example, consider there are some trucks that load products into some trains for transportation. It might be the case that adding a new train does not provide any opportunity for gaining extra benefit because the current number of trucks is too low and they cannot fill the trains in time. In this case, we can say that the number of trucks is a bottleneck. Although it is easy to define bottleneck intuitively, it is not trivial to define this term in general.

There are a few different definitions for bottlenecks. These definitions are categorized into five groups in [13]: (i) capacity based definitions, (ii) critical path based definitions, (iii) structure based definitions, (iv) algorithm based definitions, and (v) system performance based definitions. It was claimed that none of these definitions was comprehensive and some examples were provided to support this claim. Also, a new definition was proposed which was claimed to be the most comprehensive definition for a bottleneck: “a set of constraints with positive average shadow price” [13]. In fact, the average shadow price in a linear and integer linear program can be considered as a measure for bottlenecks in a system [30].

Although ASP can be useful in determining the bottlenecks in a system, it has some limitations when it comes to removing bottlenecks. In this section, we discuss some limitations of removing bottlenecks based on ASP.

Obviously, the concept of ASP has been only defined for LP and MILP, but not for problems with non-linear objective functions and constraints. Thus, using the concept of ASP prevents us from identifying and removing bottlenecks in a non-linear system.

Let us consider the following simple problem[5] (the problem is extremely simple and it has been only given as an example to clarify limitations of the previous definitions): in a mine operation, there are 19 trucks and two trains. Trucks are used to fill trains with some products and trains are used to transport products to a destination. The rate of the operation for each truck is 100 tonnes/h (tph) and the capacity of each train is 2,000 tonnes. What is the maximum tonnage that can be loaded to the trains in 1 h? The ILP model for this problem is given by:

$$\begin{gathered}\text{find } x \text{ and } y \text{ s.t. } z = \max\{2000y\} \text{ subject to} \\ g_1 : 2000y - 100x \leq 0, g_2 : x \leq 19, g_3 : y \leq 2\end{gathered} \tag{9}$$

where $x \geq 0$ is the number of trucks and $y \geq 0$ is the number of loaded trains (y can be a floating point value which refers to partially loaded trains). The constraint g_1 limits the amount of products loaded by the trucks into the trains (trucks cannot overload the trains). The solution is obviously $y = 0.95$ and $x = 0.19$ with objective value 1,900. We also calculated the value of ASP for all three constraints:

- ASP for g_1 is 1: by adding one unit to the first constraint ($2000y - 100x \leq 0$ becomes $2000y - 100x \leq 1$) the objective value increases by 1,
- ASP for g_2 is 100: by adding 1 unit to the second constraint ($x \leq 19$ becomes $x \leq 20$) the objective value increases by 100,
- ASP for g_3 is 0: by adding 1 unit to the second constraint ($y \leq 2$ becomes $y \leq 3$) the objective value does not increase.

Accordingly, the first and second constraints are bottlenecks as their corresponding ASPs are positive. Thus, it would be beneficial if investments are concentrated on adding one unit to the first or second constraint to improve the objective value.

[5] We have made several such industry-inspired stories and benchmarks available: http://cs.adelaide.edu.au/~optlog/research/bottleneck-stories.htm.

Adding one unit to the first constraint is meaningless from the practical point of view. In fact, adding one unit to RHS of the constraint g_1 means that the amount of products that is loaded into the trains can exceed the trains' capacities by one ton, which is not justifiable. In the above example, there is another option for the decision maker to achieve a better solution: if it is possible to improve the operation rate of the trucks to 101 tph, the best achievable solution is improved to 1,919 tons. Thus, it is clear that the bottleneck might be a specification of a resource (the operation rate of trucks in our example) that is expressed by a value in the coefficients matrix and not necessarily RHS.

Thus, it is clear that ASP only gives information about the impact of changing RHS in a constraint, while the bottleneck might be a value in the coefficient matrix. The commonly used ASP, which only gives information about the impact of changing RHS in a constraint, cannot identify such bottlenecks. Figure 3 illustrates this limitation.

The value of ASP represents only the effects of changing the value of RHS of the constraints (Fig. 3, left) on the objective value while it does not give any information about the effects the values in the coefficients matrix might have on the objective value (constraint g_1 in Fig. 3, right). However, as we are show in our example, it is possible to change the values in the coefficient matrix to make investments in order to remove bottlenecks.

The value of ASP does not provide any information about the best strategy of selecting bottlenecks to remove. In fact, it only provides information about the benefit of elevating the RHS in each constraint and does not say anything about the order of significance of the bottlenecks. It remains the task of the decision maker to compare different scenarios (also known as *what-if* analysis). For example, from a managerial point of view, it is important to answer the following question: is adding one unit to the first constraint (if possible) better than adding one unit to the second constraint (purchase a new truck)? Note that in real-world problems, there might be many

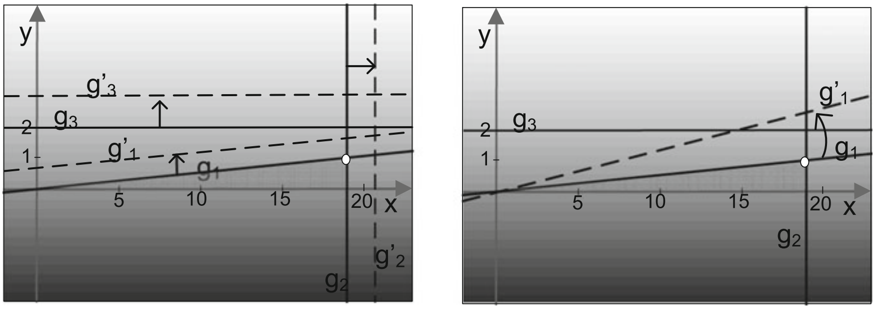


Fig. 3 x and y are number of trucks and number of trains respectively, *gray gradient* indication of objective value (the lighter the better), *shaded area* feasible area, g_1, g_2, g_3 are constraints, the white point is the best feasible point

resources and constraints, and a manual analysis of different scenarios might be prohibitively time consuming. Thus, a smart strategy is needed to find the best set of to-be-removed bottlenecks in order to gain maximum profit with lowest investment. In summary, the limitations of identifying bottlenecks using ASP are:

- **Limitation 1**: ASP is only applicable if objective and constraints are linear.
- **Limitation 2**: ASP does not evaluate changes in the coefficients matrix (the matrix A) and it is only limited to RHS.
- **Limitation 3**: ASP does not provide information about the strategy for investment in resources, and the decision maker has to manually conduct analyses to find the best investment strategy.

In order to resolve the limitations of ASP we proposed a new definition for bottlenecks and a new formulation for investment [11]. We defined bottlenecks as follows: *A bottleneck is a modifiable specification of resources that by changing its value, the best achievable performance of the system is improved*. Note that this definition is a generalization of the definition of bottleneck in [13]: a set of constraints with positive average shadow price is defined as a bottleneck. In fact, the definition in [13] concentrated on RHS only (it is just about the average shadow price) and it considers a bottleneck as a set of constraints. Conversely, our definition is based on any modifiable coefficient in the constraints (from capacity, to rates, or availability) and it introduces each specification of resources as a potential bottleneck.

Also, in order to determine the best possible investment to a system, we defined a Bottleneck COP (BCOP) for any COP as follows:

$$\text{find } x \text{ and } l \text{ s.t. } z = \begin{cases} \max f(x, l) \\ \min B(l) \end{cases} \text{ subject to } g_i(x, l_i) \le 0 \text{ for all } i \tag{10}$$

where l is a vector (l might contain continuous or discrete values) which contains l_i for all i and $B(l)$ is a function that calculates the cost of modified specifications of resources coded in the vector l. For any COP, we can define a corresponding BCOP and by solving the BCOP, the plan for investment is determined.

The identification of bottlenecks and their removal are important topics in real-world optimization. As it was mentioned earlier, locating bottlenecks and finding the best possible investment is of a great importance in large industries. For example, in the mining process described in Sect. 2 not only the number of trucks, trains, or other resources can constitute a bottleneck, but also the operation rate of any of these resources can also constitute a bottleneck. Given the expenses for removing any of these bottlenecks, one can use the model in Eq. 10 to identify the best way of investment to grow the operations and make the most benefit. This area has remained untouched by the EC community, while there are many opportunities to apply EC-based methodologies to deal with bottlenecks and investments.

6 Discussion and Future Directions

Clearly, all three research directions (multi-component problems, edge of feasibility, and bottlenecks and investment) are relevant for solving real-world problems.

First, as it was mentioned earlier, an optimal solution for each component does not guarantee global optimality, so that a solution that represents the global optimum does not necessarily contain good schedules for each component in isolation [36]. The reason lies on the dependency among components. In fact, because of dependency, even if the best solvers for each component are designed and applied to solve each component in isolation, it is not useful in many real-world cases—the whole problem with dependency should be treated without decomposition of the components. Note that, decomposing problems that are not dependent on each other can be actually valuable as it makes the problem easier to solve. However, this decomposition should be done carefully to keep the problem unchanged. Of course complexity of decomposing multi-component problems is related to the components dependencies. For example, one can define a simple dependency between KP and TSP in a TTP problem that makes the problems decomposable or make them tighten together so that they are not easily decomposable.

Looking at dependencies among components, the lack of abstract problems that reflect this characteristic is obvious in the current benchmarks. In fact, real-world supply chain optimization problems are a combination of many smaller sub-problems dependent on each other in a network while benchmark problems are singular. Because global optimality is in interest in multi-component problems, singular benchmark problems cannot assess quality of methods which are going to be used for multi-component real-world problems with the presence of dependency.

Multi-component problems pose new challenges for the theoretical investigations of evolutionary computation methods. The computational complexity analysis of evolutionary computation is playing a major role in this field [2, 42]. Results have been obtained for many NP-hard combinatorial optimization problems from the areas of covering, cutting, scheduling, and packing. We expect that the computational complexity analysis can provide new rigorous insights into the interactions between different components of multi-component problems. As an example, we consider again the TTP problem. Computational complexity results for the two underlying problems (KP and TSP) have been obtained in recent years. Building on these results, the computational complexity analysis can help to understand when the interactions between KP and TSP make the optimization process harder.

Second, there has been some experimental evidence that showed the importance of searching the boundaries of feasible and infeasible areas in a constraint optimization problem (COP) [40, 49, 50]. This boundary is defined as: the points that are feasible and the value of *at least one* of the constraints is zero for them. In [5] three new instances (called Constraint Boundary Narrower, CBN, Subset CBN, SCBN, and All in a subset CBN, ACBN) for the constraint violation function were proposed which were able to reduce the feasible area to only boundaries of the feasible area. In the SCBN (ACBN), it is possible to select a subset of constraints and limit the boundaries where *at least one* of these constraints (*all* of these constraints) is (are) active. The thickness of the boundaries was adjustable in the proposed method by a parameter (ϵ). Experiments showed that changing the value of ϵ influences the performance of the algorithm. In fact, a smaller value of ϵ causes limiting the feasible area to narrower boundaries, which makes finding the feasible areas harder. However, although it is harder to find the feasible areas (narrower boundaries), improving the final solutions is easier once the correct boundary was found. Thus, as a potential future work, one can design an adaptive method so that the search begins by exploring the feasible area and later concentrates on the boundaries.

Finally, a new definition for bottlenecks and a new model to guide decision makers to make the most profitable investment on their system should assist in narrowing the gap between what is being considered in academia and industry. Our definition for bottlenecks and model for investment overcomes several of the drawbacks of the model that is based on average shadow prices:

- It can work with non-linear constraints and objectives.
- It offers changes to the coefficient matrix.
- It can provide a guide towards optimal investments.

This more general model can form the basis for more comprehensive analytical tools as well as improved optimization algorithms. In particular for the latter application, we conjecture that nature-inspired approaches are adequate, due to the multi-objective formulation of the problem and its non-linearity.

Bottlenecks are ubiquitous and companies make significant efforts to eliminate them to the best extent possible. To the best of our knowledge, however, there seems to be very little published research on approaches to identify bottlenecks research on optimal investment strategies in the presence of bottlenecks seems to be even non-existent. In the future, we will push this research further, in order to improve decision support systems. If bottlenecks can be identified efficiently, then this information can be easily shown to the decision maker, who can then subsequently use this information in a manual optimization process.

There is also another research direction recently introduced to address real-world optimization problems that is locating disjoint feasible regions in a search space [4, 8].[6] It has been argued that the feasible area in constrained optimization problems might have an irregular shape and might contain many disjoint regions. Thus, it is beneficial if an optimization algorithm can locate these regions as much as possible so that the probability of finding the region that contain the best feasible solution is increased. The problem of locating many disjoint feasible regions can be viewed as niching in multi-modal optimization [4].

[6] we have excluded this topic from this chapter because of the lack of space.

References

1. Ackoff RL (1979) The future of operational research is past. J Oper Res Soc 53(3):93–104. ISSN 0160–5682
2. Auger A, Doerr B (2011) Theory of randomized search heuristics: foundations and recent developments, vol 1. World Scientific. ISBN 9814282669
3. Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. SIAM Rev 53(3):464–501. ISSN 0036–1445
4. Bonyadi MR, Michalewicz Z (2014) Locating potentially disjoint feasible regions of a search space with a particle swarm optimizer, book section to appear. Springer, New York
5. Bonyadi MR, Michalewicz Z (2014) On the edge of feasibility: a case study of the particle swarm optimizer. In: Congress on evolutionary computation, IEEE, pp 3059–3066
6. Bonyadi MR, Li X, Michalewicz Z (2013) A hybrid particle swarm with velocity mutation for constraint optimization problems. In: Genetic and evolutionary computation conference, ACM, pp 1–8. doi:10.1145/2463372.2463378
7. Bonyadi MR, Michalewicz Z, Barone L (2013) The travelling thief problem: the first step in the transition from theoretical problems to realistic problems. In: Congress on evolutionary computation, IEEE
8. Bonyadi MR, Li X, Michalewicz Z (2014) A hybrid particle swarm with a time-adaptive topology for constrained optimization. Swarm Evol Comput 18:22–37. doi:10.1016/j.swevo.2014.06.001
9. Bonyadi MR, Michalewicz Z, Neumann F, Wagner M (2014) Evolutionary computation for multi-component problems: opportunities and future directions. Frontiers in Robotics and AI, Computational Intelligence, under review, 2014
10. Bonyadi MR, Michalewicz Z, Przybyek MR, Wierzbicki A (2014) Socially inspired algorithms for the travelling thief problem. In: Genetic and evolutionary computation conference (GECCO), ACM
11. Bonyadi MR, Michalewicz Z, Wagner M (2014) Beyond the edge of feasibility: analysis of bottlenecks. In: International conference on simulated evolution and learning (SEAL), volume To appear, Springer
12. Charnes A, Cooper WW (1957) Management models and industrial applications of linear programming. Manag Sci 4(1):38–91. ISSN 0025–1909
13. Chatterjee A, Mukherjee S (2006) Unified concept of bottleneck. Report, Indian Institute of Management Ahmedabad, Research and Publication Department
14. Cho S, Kim S (1992) Average shadow prices in mathematical programming. J Optim Theory Appl 74(1):57–74
15. Crema A (1995) Average shadow price in a mixed integer linear programming problem. Eur J Oper Res 85(3):625–635. ISSN 0377–2217
16. Frieze A (1975) Bottleneck linear programming. Oper Res Q 26(4):871–874
17. Goldratt EM (1990) Theory of constraints. North River, Croton-on-Hudson
18. Goldratt EM, Cox J (1993) The goal: a process of ongoing improvement. Gower, Aldershot
19. Heywood MI, Lichodzijewski P (2010) Symbiogenesis as a mechanism for building complex adaptive systems: a review. In: Applications of evolutionary computation, Springer, pp 51–60
20. Hillis WD (1990) Co-evolving parasites improve simulated evolution as an optimization procedure. Phys D: Nonlinear Phenom 42(1):228–234. ISSN 0167–2789
21. Jacob Stolk AMZM, Mann I (2013) Combining vehicle routing and packing for optimal delivery schedules of water tanks. OR Insight 26(3):167190. doi:10.1057/ori.2013.1
22. Jin Y, Branke J (2005) Evolutionary optimization in uncertain environments-a survey. IEEE Trans Evol Comput 9(3):303–317. ISSN 1089–778X
23. Keane A (1994) Genetic algoritm digest. ftp://ftp.cse.msu.edu/pub/GA/gadigest/v8n16.txt
24. Keen PG (1981) Value analysis: justifying decision support systems. MIS Q 5:1–15. ISSN 0276–7783
25. Kim S, Cho S-C (1988) A shadow price in integer programming for management decision. Eur J Oper Res 37(3):328–335. ISSN 0377–2217

26. Koopmans TC (1977) Concepts of optimality and their uses. Am Econ Rev 67:261–274. ISSN 0002–8282
27. Lau HC, Song Y (2002) Combining two heuristics to solve a supply chain optimization problem. Eur Conf Artif Intell 15:581–585
28. Leguizamon G, Coello CAC (2009) Boundary search for constrained numerical optimization problems with an algorithm inspired by the ant colony metaphor. IEEE Trans Evol Comput 13(2):350–368. ISSN 1089–778X
29. Li X, Bonyadi MR, Michalewicz Z, Barone L (2013) Solving a real-world wheat blending problem using a hybrid evolutionary algorithm. In: Congress on evolutionary computation, IEEE, pp 2665–2671. ISBN 1479904538
30. Luebbe R, Finch B (1992) Theory of constraints and linear programming: a comparison. Int J Prod Res 30(6):1471–1478. ISSN 0020–7543
31. Maksud Ibrahimov SSZM, Mohais A (2012) Evolutionary approaches for supply chain optimisation part 1. Int J Intell Comput Cybern 5(4):444–472
32. Maksud Ibrahimov SSZM, Mohais A (2012) Evolutionary approaches for supply chain optimisation part 2. Int J Intell Comput Cybern 5(4):473–499
33. Martello S, Toth P (1990) Knapsack problems: algorithms and computer implementations. Wiley, Chichester
34. Mersmann O, Bischl B, Trautmann H, Wagner M, Bossek J, Neumann F (2013) A novel feature-based approach to characterize algorithm performance for the traveling salesperson problem. Ann Math Artif Intell 1–32. ISSN 1012–2443
35. Michalewicz Z (1992) Genetic algorithms + data structures = evolution programs. Springer. ISBN 3540606769
36. Michalewicz Z (2012) Quo vadis, evolutionary computation? Adv Comput Intell 98–121
37. Michalewicz Z (2012) Ubiquity symposium: evolutionary computation and the processes of life: the emperor is naked: evolutionary algorithms for real-world applications. Ubiquity, 2012(November):3
38. Michalewicz Z, Fogel D (2004) How to solve it: modern heuristics. Springer, New York. ISBN 3540224947
39. Michalewicz Z, Schoenauer M (1996) Evolutionary algorithms for constrained parameter optimization problems. Evol Comput 4(1):1–32. ISSN 1063–6560
40. Michalewicz Z, Nazhiyath G, Michalewicz M (1996) A note on usefulness of geometrical crossover for numerical optimization problems. In: Fifth annual conference on evolutionary programming, Citeseer, p 305312
41. Nallaperuma S, Wagner M, Neumann F, Bischl B, Mersmann O, Trautmann H (2013) A feature-based comparison of local search and the christofides algorithm for the travelling salesperson problem. In: Proceedings of the twelfth workshop on foundations of genetic algorithms XII, ACM, pp 147–160. ISBN 1450319904
42. Neumann F, Witt C (2012) Bioinspired computation in combinatorial optimization: algorithms and their computational complexity. In: Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference companion, ACM, pp 1035–1058. ISBN 1450311784
43. Nguyen T, Yao X (2012) Continuous dynamic constrained optimisation-the challenges. IEEE Trans Evol Comput 16(6):769–786. ISSN 1089–778X
44. Polyakovskiy S, Bonyadi MR, Wagner M, Michalewicz Z, Neumann F (2014) A comprehensive benchmark set and heuristics for the travelling thief problem. In: Genetic and evolutionary computation conference (GECCO), ACM. ISBN 978-1-4503-2662-9/14/07. doi:10.1145/2576768.2598249
45. Potter M, De Jong K (1994) A cooperative coevolutionary approach to function optimization. In: Parallel problem solving from nature, Springer, Berlin Heidelberg, pp 249–257. doi:10.1007/3-540-58484-6269
46. Rahman S-U (1998) Theory of constraints: a review of the philosophy and its applications. Int J Oper Prod Manage 18(4):336–355. ISSN 0144–3577

47. Rosin CD, Belew RK (1995) Methods for competitive co-evolution: finding opponents worth beating. In: ICGA, pp 373–381
48. Runarsson T, Yao X (2000) Stochastic ranking for constrained evolutionary optimization. IEEE Trans Evol Comput 4(3):284–294. ISSN 1089–778X
49. Schoenauer M, Michalewicz Z (1996) Evolutionary computation at the edge of feasibility. In: Parallel problem solving from nature PPSN IV, pp 245–254
50. Schoenauer M, Michalewicz Z (1997) Boundary operators for constrained parameter optimization problems. In: ICGA, pp 322–32
51. Smith-Miles K, van Hemert J, Lim XY (2010) Understanding TSP difficulty by learning from evolved instances, Springer, pp 266–280. ISBN 3642137997
52. Smith-Miles K, Baatar D, Wreford B, Lewis R (2014) Towards objective measures of algorithm performance across instance space. Comput Oper Res 45:12–24. ISSN 0305–0548
53. Weise T, Zapf M, Chiong R, Nebro A (2009) Why is optimization difficult? Nature-inspired algorithms for optimisation, pp 1–50
54. Wu ZY, Simpson AR (2002) A self-adaptive boundary search genetic algorithm and its application to water distribution systems. J Hydraul Res 40(2):191–203. ISSN 0022–1686

Selection of Significant Features Using Monte Carlo Feature Selection

Susanne Bornelöv and Jan Komorowski

Abstract Feature selection methods identify subsets of features in large datasets. Such methods have become popular in data-intensive areas, and performing feature selection prior to model construction may reduce the computational cost and improve the model quality. Monte Carlo Feature Selection (MCFS) is a feature selection method aimed at finding features to use for classification. Here we suggest a strategy using a z-test to compute the significance of a feature using MCFS. We have used simulated data with both informative and random features, and compared the z-test with a permutation test and a test implemented into the MCFS software. The z-test had a higher agreement with the permutation test compared with the built-in test. Furthermore, it avoided a bias related to the distribution of feature values that may have affected the built-in test. In conclusion, the suggested method has the potential to improve feature selection using MCFS.

Keywords Feature selection · MCFS · Monte Carlo · Feature significance · Classification

1 Introduction

With the growth of large datasets in areas such as bioinformatics, computational chemistry, and text recognition, limitations in the computational resources may force us to restrict the analysis to a subset of the data. Feature selection methods reduce the

S. Bornelöv · J. Komorowski (✉)
Department of Cell and Molecular Biology, Science for Life Laboratory,
Uppsala University, Uppsala, Sweden
e-mail: jan.komorowski@icm.uu.se

S. Bornelöv
Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden
e-mail: susanne.bornelov@imbim.uu.se

J. Komorowski
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_2

data by selecting a subset of the features. An assumption in feature selection is that large datasets contain some redundant or non-informative features. If successfully removing those, both the speed of the model training, the performance, and the interpretation of the model may be improved [1].

There are several feature selection methods available. For a review of feature selection techniques used in bioinformatics, see Saeys et al. [2]. Some methods are univariate and consider one feature at a time; others include feature interactions to various degrees. In this paper we have studied Monte Carlo Feature Selection (MCFS) [3]. MCFS focuses on selecting features to be used for classification. The use of MCFS was originally illustrated by selecting genes with importance for leukemia and lymphoma [3], and it was later used to study e.g. HIV-1 by selecting residues in the amino acid sequence of reverse transcriptase with importance for drug resistance [4, 5]. Furthermore, MCFS may be used to rank the features based on their relative importance score. Thus, MCFS may be applied even on smaller datasets if the aim is to rank the features by their impact on the outcome (see e.g. [6–8]).

MCFS is a multivariate feature selection method based on random sampling of the original features. Each sample is used to construct a number of decision trees. Each feature is then given a score—relative importance (RI)—according to how it performs in the decision trees. Thus, the selection of a feature is explicitly based on how the feature contributes to classification.

One question is how to efficiently interpret the RI of a feature. If MCFS is used to select a subset suitable for classification, a strategy may be to select the x highest ranked features [6]. However, a stronger statistical basis for making the cutoff would be preferred, particularly, when MCFS is used to determine which features significantly influence the outcome.

The MCFS algorithm is implemented in the dmLab software available at [9]. There is a statistical test on the significance of a feature implemented in the software. The strategy of the test is to perform a number of permutations of the decision column, and in each permutation save the highest RI observed for any feature. Thereafter, the test compares the RI of each feature in the original data to the 95 % confidence interval of the mean of the best RI scores [5].

Here, we suggest a different methodology that tests each feature separately to its own set of controls. We show that this methodology leads to more accurate results and allows us to identify the most significant feature even when they do not have the highest RI. Furthermore, by testing each feature separately, we avoid biases related to the distribution of feature values. Our suggested methodology is supported by experiments using simulated data.

In conclusion, we have provided a methodology for computing the significance of a feature using MCFS. We have shown that this methodology improves the currently used statistical test, and discussed the implications of using alternative methods.

2 Materials and Methods

2.1 Monte Carlo Feature Selection

The MCFS algorithm is based on extensive use of decision trees. The general idea is to select s subsets of the original d features, each with a random selection of m features. Each such subset is divided into a training and test set with 2/3 and 1/3 of the objects, respectively. This division is repeated t times, and a decision tree classifier is trained on each training set. In all, st decision trees are trained and evaluated on their respective test set. An overview of the methodology is shown in Fig. 1.

Each feature is scored according to how it performs in these classifiers by a score called relative importance (RI). The RI of a feature g was defined by Draminski et al. [3] as

$$RI_g = \frac{1}{M_g} \sum_{\tau=1}^{st} (wAcc_\tau)^u \sum_{n_g(\tau)} \mathrm{IG}(n_g(\tau)) \left(\frac{\text{no.in}\, n_g(\tau)}{\text{no.in}\, \tau} \right)^v \tag{1}$$

where s is the number of subsets and t is the number of splits for each subset. M_g is the number of times the attribute g was present in the training set used to construct a decision tree. For each tree τ the weighted accuracy $wAcc$ is calculated as the mean sensitivity over all decision classes, using

$$wAcc = \frac{1}{c} \sum_{i=1}^{c} \frac{n_{ii}}{n_{i1} + n_{i2} + \cdots + n_{ic}} \tag{2}$$

where c is the number of decision classes and n_{ij} is the number of objects from class i that were classified to class j.

Furthermore, for each $n_g(\tau)$ (a node n in decision tree τ that uses attribute g) the information gain (IG) of $n_g(\tau)$ and the fraction of the number of training set objects in (no.in) $n_g(\tau)$ compared to the number of objects in the tree root is computed. There are two weighting factors u and v that determine the importance of the $wAcc$ and the number of objects in the node.

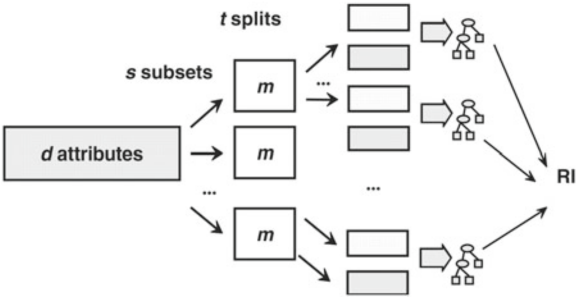


Fig. 1 Overview of the MCFS procedure. Reproduced from Draminski et al. [3]

2.2 Construction of Datasets

To apply MCFS and to compute the significance of the features, we constructed datasets with 120 numerical and 120 binary features. For each type of features, 20 were correlated to the decision and 100 were uncorrelated. The decision class was defined to be binary (0 or 1) with equal frequency of both decisions. The number of simulated objects was set to either 100 or 1,000. Thus, for each object the decision class value was randomly drawn from the discrete uniform distribution [0,1] prior to generating the attribute values. Detailed description of the attributes is provided in the following sections. To verify that the features with an expected correlation to the decision indeed were correlated, the Pearson correlation between each non-random feature and the decision was computed after the data generation (Table 1).

Numerical Uncorrelated Features: *RandNum*$_0$ to *RandNum*$_{99}$. The values of a numerical uncorrelated feature ($RandNum_i$, $0 \le i \le 99$) were randomly drawn from the discrete uniform distribution [1, $i+1$]. Thus, the indices defined the range of

Table 1 Pearson correlation between each correlated feature and the decision. Presented for both datasets (100 objects and 1,000 objects) separately

i	100 objects		1,000 objects	
	Num_i	Bin_i	Num_i	Bin_i
0	0.74	0.96	0.74	0.95
1	0.72	0.94	0.65	0.91
2	0.58	0.86	0.63	0.87
3	0.66	0.84	0.50	0.81
4	0.50	0.77	0.50	0.78
5	0.53	0.73	0.47	0.69
6	0.19	0.60	0.43	0.66
7	0.39	0.64	0.41	0.64
8	0.34	0.56	0.35	0.60
9	0.28	0.54	0.35	0.55
10	0.38	0.39	0.28	0.46
11	0.22	0.41	0.29	0.41
12	0.18	0.33	0.23	0.45
13	0.21	0.30	0.20	0.31
14	0.29	0.33	0.14	0.32
15	0.18	0.19	0.16	0.32
16	0.15	0.31	0.16	0.18
17	–0.01	0.01	0.07	0.14
18	0.08	0.07	0.07	0.15
19	–0.06	–0.02	–0.03	0.05

possible values, which allowed us to test whether the number of possible values for a feature influenced its ranking.

Numerical Correlated Features: *Num_0* to *Num_{19}*. The values of a numerical correlated feature (Num_i, $0 \leq i \leq 19$) were defined using the following algorithm: Let X be a random variable from the continuous uniform distribution (0,1). If $X > (i+1)/21$ the value was selected randomly from the binomial distribution $B(6, 0.5)$ if $Decision = 0$, and from $B(6, 0.5) + 3$ if $Decision = 1$. Otherwise, if $X \leq (i+1)/21$, the value was selected randomly from the uniform distribution [0, 9]. Thus, low values were indicative of $Decision = 0$ and high values of $Decision = 1$, with a noise level indicated by the feature index.

Binary Uncorrelated Features: *$RandBin_0$* to *$RandBin_{99}$*. The values of a binary uncorrelated feature ($RandBin_i$, $0 \leq i \leq 99$) were defined using the following algorithm: Let X be a random variable from the continuous uniform distribution (0,1). If $X > (i+1)/101$ the value is 1, otherwise it is 0.

Thus, features with low indices will have ones in excess, features with middle indices will have more even distribution of ones and zeroes, and those with high indices will have zeroes in excess.

Binary Correlated Features: *Bin_0* to *Bin_{19}*. The values of a binary correlated feature (Bin_i, $0 \leq i \leq 19$) were defined using the following algorithm: Let X_1 be a random variable from the continuous uniform distribution (0,1). If $X_1 > (i+1)/21$, the value is equal to the decision. Otherwise it is assigned by drawing another random variable X_2 from the continuous uniform distribution (0,1). If $X_2 > (i+1)/21$, the value is 1, otherwise it is 0.

2.3 Performing the Experiments

The experiments were performed using the dmLab software version 1.85. We applied the rule-of-thumb to set the number of features selected in each subset to $\sqrt{d}$, where d is the total number of features. Thus using 240 features, we used $m = \sqrt{240} \approx 15$. The number of subsets was set to s = 3,000 for the permutation runs and s = 100,000 for the original data. The number of trees trained in each subset was set to t = 5 and the number of permutation test runs was set to *cutPointRuns* = 10,000. The weighting parameters were set to $u = 0$ and $v = 1$.

There were two main arguments for using a higher number of subsets on the original data. Firstly, ranking of the features in the original data is the most crucial part of the experiment. Therefore, it is generally motivated to focus more of the computational resources onto this step. Secondly, both the z-test and the built-in test require the rankings of the original data to be stable, which is obtained by constructing a high number of subsets.

Setting $u = 0$ will omit the decision tree accuracy from the calculation of RIs. Indeed, using model performance as a selection criteria may be counter-productive

[10], and our experience is that the inclusion of the accuracy in the calculation of the RI overestimates the importance of all features in the original data compared to the permuted ones. This effect is expected, since the accuracy on the original data will reflect the most predictive features, whereas on the permuted data it will only reflect random variation of the decision trees.

2.4 Selection of Significant Features

In this section we present different strategies to estimate the p-value of the RI of a feature using a permutation test, either alone or in combination with additional tests. Using a traditional permutation test requires thousands of permutations to yield efficient estimates of small p-values. Thus, alternative tests performing a smaller number of permutations and using these to estimate the underlying distribution may save computational time. The test that is built-in into dmLab employs this strategy and performs a t-test comparing the best RIs obtained during the permutation runs to the RI of a feature on the original data. Here we suggest another approach using a z-test to compute the p-value by estimating a normal distribution for each feature separately.

During the permutation test the number of permutations, N, was set to 10,000 to obtain sufficient resolution of the p-values. The permutation test p-values were then used as a gold standard to evaluate the build-in test and the suggested z-test. For these tests a substantially smaller number of permutations are needed. Consequently, we used only the 100 first permutation runs to estimate the p-values using the built-in and the z-test.

Using a Permutation Test to Select Significant Features. A permutation test may be applied to compute an approximation of the empirical p-value of a RI. The null hypothesis is that the RI calculated on the real data is no better than the RIs computed for the permutated data. The empirical p-value approximates the probability of observing a test statistics at least as extreme as the observed value, assuming that the null hypothesis is true. Typically, a significance level, such as 0.05, is defined and attributes associated with p-values below this level are considered significantly informative.

Theoretically, the true permutation test p-value of RI = x that was measured for a feature g would be

$$p_{true}(RI_g = x) = \frac{\sum_{i=1}^{N_{all}} \mathbf{I}(RI_g^i \geq x)}{N_{all}} \tag{3}$$

where **I** is the indicator function taking value 1 if the condition is met, and 0 otherwise. $RI^i{}_g$ is the RI of the attribute g in permutation $i$and N_{all} denotes the total number of possible permutations. However, since N_{all} may be extremely large, only a limited number of permutations are commonly performed. Furthermore, pseudo-counts are added to avoid p-values of zero, which are theoretically impossible since at least one possible permutation has to be identical to the original data. Thus, an approximation of the permutation test p-value is commonly applied, which is based on the N number of permutations with $N \ll N_{all}$ using the following expression

$$p(RI_g = x) = \frac{1 + \sum_{i=1}^{N} \mathbf{I}(RI_g^i \geq x)}{N + 1} \tag{4}$$

Using a z-Test to Select Significant Features. By performing N permutations, each feature receives N estimates of its relative importance on non-informative data. If $N > 30$ and the RIs are normally distributed, the distribution mean μ_g and standard deviation σ_g of a feature g may be estimated from the data as

$$\mu_g = \frac{1}{N} \sum_{i=1}^{N} RI_g^i \tag{5}$$

and

$$\sigma_g = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (RI_g^i - \mu_g)^2} \tag{6}$$

where $RI^i{}_g$ is the RI of attribute g in permutation i.

Thus, the z-score of the RI for a feature g on the original data, $RI_g = x$, may be computed as

$$z = (x - \mu_g)/\sigma_g. \tag{7}$$

A z-test can be applied to calculate the p-value associated to a particular z-score. Since no feature is expected to perform significantly worse on the original data compared with the permuted one, an upper-tail p-value was computed.

Using the Built-in Test to Select Significant Features. To compare our results, we also used the combined permutation test implemented in the dmLab software. This test is also based on N permutations of the decision, and using each such permuted dataset, the whole MCFS procedure is repeated and the RI of each feature is computed. As opposed to the previous strategies, only the highest RI from each permuted dataset ($RI_{\max}$) is used, independently of which feature it is based on. Thus, N such $RI_{\max}$ values are generated and used to estimate the parameters $\mu_{\max}$ and $\sigma_{\max}$ applying

$$\mu_{\max} = \frac{1}{N} \sum_{i=1}^{N} RI_{\max}^i \tag{8}$$

and

$$\sigma_{\max} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(RI^{i}_{\max} - \mu_{\max})^{2}}. \tag{9}$$

A t-statistic is then computed per feature g as

$$T = (x_g - \mu_{\max})/(\sigma_{\max}/\sqrt{N}) \tag{10}$$

and the two-sided p-value associated to the t-statistics is obtained.

3 Results

3.1 Results of Simulation Study

We applied MCFS to the datasets with 100 and 1,000 objects. Table 2 summarizes the results after MCFS using 100 objects. The RI of each feature is reported, as well as the estimated RI mean and standard deviation on the permuted data. The 10,000 RIs computed for each feature on the permuted data were approximately bell shaped, occasionally displaying a bias towards either of the distribution tails. The p-values were calculated using the z-test and the permutation test as described in Sect. 2.4. Additionally, an overall RI threshold at the 0.05 significance level was estimated to 0.0787 using the built-in method in dmLab.

Using both the z-test and the permutation test Num_0-Num_5, Num_7-Num_8, Num_{10}, Bin_0- Bin_{11}, Bin_{14}, and Bin_{16} were significant at the 0.05 level. Using the built-in t-test combining all features, the Bin_{10}-Bin_{11}, Bin_{14}, and Bin_{16} were not identified as significant since their RI was below 0.0787. Note that Bin_{16} was significant according to the z-test and the permutation test, although it had a lower RI than Num_9 that was not significant using any of the tests.

A notable association between the ranking of the random binary features and their indices was observed (Fig. 2a), where features with intermediate indices were ranked higher than those with low or high indices. Since the random binary features with low or high indices were defined to have an excess of ones or zeroes, respectively, this corresponds to a weak preference for features with a uniform distribution of values. However, no relation between the value range of a feature and its relative importance was observed, consistent with previously reported results [11], although the variation of the RIs increased slightly with the value range (Fig. 2b). Both the binary and numeric features were scored according to their expected relevance (Fig. 2c, d).

Since the data was randomly generated simulating only 100 objects, the exact size of the effect for a feature may differ slightly from the expectation. Thus, we repeated the same methodology with a sample size of 1,000 objects instead. The results are shown in Table 3. This time, Bin_0-Bin_{15} and Num_0-Num_{13} were selected using z-test

Table 2 Results of MCFS on simulated data with 100 objects. Significant features and the three highest ranked non-significant features of each type are shown. The features are ranked according to their RI. Grayed lines denote non-significant features

Rank	Feature	RI	μ_{RIperm}	σ_{RIperm}	$p_{\text{z-test}}$	$p_{\text{perm-test}}$
1	Bin_0	0.859	0.0216	0.00458	<0.0001	0.0001
2	Bin_1	0.806	0.0228	0.00600	<0.0001	0.0001
3	Bin_2	0.557	0.0234	0.00548	<0.0001	0.0001
4	Bin_3	0.521	0.0240	0.00519	<0.0001	0.0001
5	Num_0	0.408	0.0344	0.00928	<0.0001	0.0001
6	Bin_4	0.398	0.0248	0.01121	<0.0001	0.0001
7	Bin_5	0.351	0.0239	0.00547	<0.0001	0.0001
8	Num_1	0.317	0.0326	0.00694	<0.0001	0.0001
9	Num_3	0.239	0.0334	0.00820	<0.0001	0.0001
10	Bin_7	0.219	0.0251	0.00438	<0.0001	0.0001
11	Num_2	0.212	0.0347	0.00941	<0.0001	0.0001
12	Num_5	0.206	0.0344	0.01061	<0.0001	0.0001
13	Bin_6	0.178	0.0249	0.00496	<0.0001	0.0001
14	Num_4	0.174	0.0342	0.00715	<0.0001	0.0001
15	Bin_9	0.146	0.0262	0.00462	<0.0001	0.0001
16	Bin_8	0.142	0.0257	0.00660	<0.0001	0.0001
17	Num_7	0.096	0.0343	0.00683	<0.0001	0.0019
18	Num_{10}	0.084	0.0357	0.01153	<0.0001	0.0024
19	Num_8	0.083	0.0353	0.01017	<0.0001	0.0032
20	Bin_{11}	0.065	0.0271	0.00733	<0.0001	0.0020
21	Bin_{10}	0.058	0.0277	0.00950	0.0006	0.0062
22	Bin_{14}	0.052	0.0229	0.00486	<0.0001	0.0127
23	Num_9	0.044	0.0361	0.00925	0.2058	0.1349
24	Bin_{16}	0.040	0.0247	0.00546	0.0032	0.0418
25	Bin_{12}	0.039	0.0274	0.00760	0.0572	0.0458
26	Num_{12}	0.038	0.0343	0.00806	0.3310	0.2664
27	Num_{13}	0.037	0.0336	0.00719	0.3127	0.2540
29	$RandNum_{73}$	0.036	0.0326	0.00632	0.2990	0.1952
31	Bin_{13}	0.034	0.0255	0.00784	0.1379	0.0829
32	$RandNum_{75}$	0.033	0.0323	0.00645	0.4753	0.3798
33	$RandBin_{31}$	0.031	0.0237	0.00647	0.1254	0.1010
34	$RandNum_{32}$	0.028	0.0299	0.00605	0.6053	0.5193
39	$RandBin_{83}$	0.022	0.0176	0.00522	0.1882	0.1670
41	$RandBin_{54}$	0.021	0.0280	0.00857	0.7759	0.9115
64	Bin_{15}	0.018	0.0244	0.01166	0.7145	0.9874

and permutation test. The threshold using the built-in test was 0.0352, which in this case identified the same features.

The relation between the RI scores and the indices of the features is shown in Fig. 3. There is a substantial decrease in the noise compared with using 100 objects.

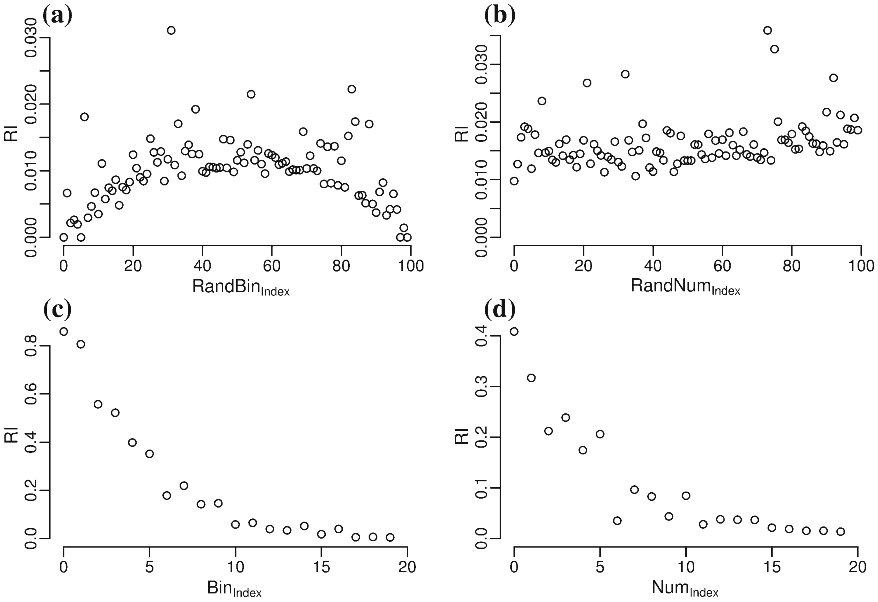


Fig. 2 Relation between attribute indices and RI using a dataset with 100 objects. Shown for **a**, **b** random and **c**, **d** informative features of both **a**, **c** binary and **b**, **d** numeric type. Note that the y-axis scale varies from panel to panel

3.2 Comparison of p-Values

In order to determine how accurate the p-values obtained through the z-test were, we compared them with the permutation test p-values (Fig. 4a). Furthermore, we computed p-values based on the built-in method, and compared to the permutation test p-values (Fig. 4b).

The p-values estimated using the z-test were closely following the ones obtained by permutation test, whereas the built-in method failed to efficiently model the empirical p-values, although the built-in method identified almost as many significant features as the z-test. Essentially, the p-values obtained by applying the built-in method were always equal to either 0 or 1. We speculate that the assumption of comparing two means results in a biased downward estimate of the variance of the data.

4 Discussion

We have used simulated data to evaluate the application of a z-test to identifying features significant for classification using MCFS. The data was designed in such

Table 3 Results of MCFS on simulated data with 1,000 objects. Significant features and the three highest ranked non-significant features of each type are shown. The features are ranked according to their RI. Grayed lines denote non-significant features

Rank	Feature	RI	μ_{RIperm}	σ_{RIperm}	$p_{\text{z-test}}$	$p_{\text{perm-test}}$
1	Bin_{0}	0.855	0.0130	0.00126	<0.0001	0.0001
2	Bin_{1}	0.732	0.0133	0.00134	<0.0001	0.0001
3	Bin_{2}	0.612	0.0136	0.00139	<0.0001	0.0001
4	Bin_{3}	0.474	0.0139	0.00149	<0.0001	0.0001
5	Num_{0}	0.439	0.0316	0.00129	<0.0001	0.0001
6	Bin_{4}	0.423	0.0139	0.00150	<0.0001	0.0001
7	Num_{1}	0.298	0.0316	0.00122	<0.0001	0.0001
8	Bin_{5}	0.290	0.0140	0.00162	<0.0001	0.0001
9	Num_{2}	0.288	0.0316	0.00139	<0.0001	0.0001
10	Bin_{6}	0.256	0.0147	0.00131	<0.0001	0.0001
11	Bin_{7}	0.224	0.0147	0.00160	<0.0001	0.0001
12	Num_{3}	0.183	0.0319	0.00149	<0.0001	0.0001
13	Bin_{8}	0.180	0.0147	0.00155	<0.0001	0.0001
14	Num_{4}	0.178	0.0320	0.00128	<0.0001	0.0001
15	Num_{5}	0.158	0.0318	0.00108	<0.0001	0.0001
16	Bin_{9}	0.152	0.0151	0.00141	<0.0001	0.0001
17	Num_{6}	0.135	0.0317	0.00129	<0.0001	0.0001
18	Num_{7}	0.127	0.0315	0.00134	<0.0001	0.0001
19	Bin_{10}	0.095	0.0151	0.00165	<0.0001	0.0001
20	Num_{8}	0.091	0.0316	0.00136	<0.0001	0.0001
21	Num_{9}	0.090	0.0316	0.00147	<0.0001	0.0001
22	Bin_{12}	0.090	0.0149	0.00145	<0.0001	0.0001
23	Bin_{11}	0.067	0.0155	0.00141	<0.0001	0.0001
24	Num_{11}	0.061	0.0312	0.00130	<0.0001	0.0001
25	Num_{10}	0.056	0.0316	0.00153	<0.0001	0.0001
26	Bin_{15}	0.044	0.0135	0.00134	<0.0001	0.0001
27	Num_{12}	0.041	0.0314	0.00153	<0.0001	0.0001
28	Bin_{14}	0.039	0.0145	0.00153	<0.0001	0.0001
29	Bin_{13}	0.037	0.0153	0.00139	<0.0001	0.0001
30	Num_{13}	0.036	0.0312	0.00149	0.0006	0.0046
31	Num_{15}	0.025	0.0311	0.00131	1.0000	1.0000
32	Num_{16}	0.025	0.0308	0.00115	1.0000	1.0000
33	Num_{14}	0.025	0.0309	0.00118	1.0000	1.0000
35	$RandNum_{7}$	0.014	0.0314	0.00134	1.0000	1.0000
36	$RandNum_{3}$	0.013	0.0317	0.00077	1.0000	1.0000
38	$RandNum_{5}$	0.013	0.0321	0.00103	1.0000	1.0000
40	Bin_{16}	0.013	0.0125	0.00137	0.3742	0.4528
137	Bin_{18}	0.010	0.0088	0.00089	0.1607	0.1239
138	Bin_{17}	0.010	0.0106	0.00135	0.7737	0.8153
140	$RandBin_{45}$	0.008	0.0156	0.00156	1.0000	1.0000
141	$RandBin_{58}$	0.008	0.0150	0.00157	1.0000	1.0000
142	$RandBin_{50}$	0.008	0.0157	0.00165	1.0000	1.0000

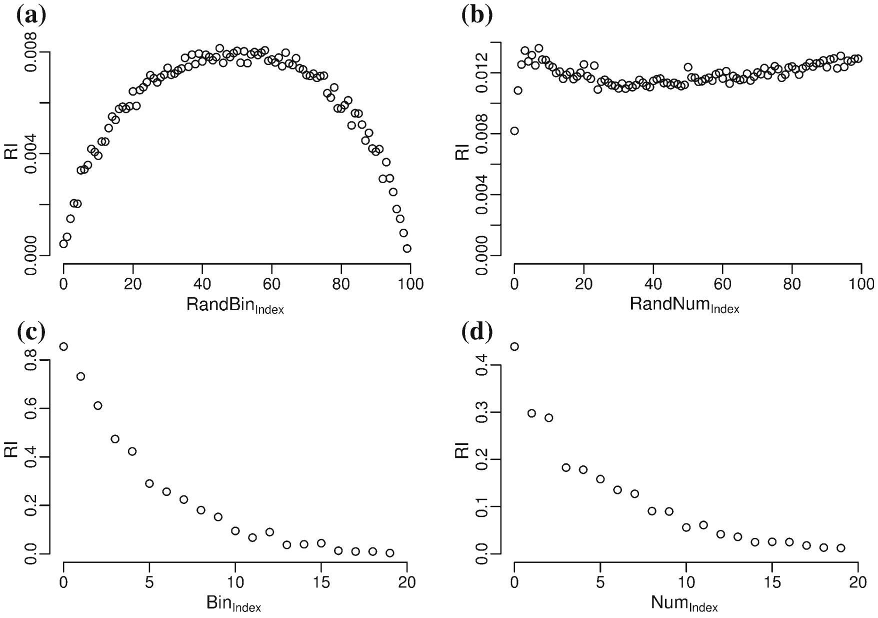


Fig. 3 Relation between attribute indices and RI using a dataset with 1,000 objects. Shown for **a**, **b** random and **c**, **d** informative features of both **a**, **c** binary and **b**, **d** numeric type. Note that the y-axis scale varies from panel to panel

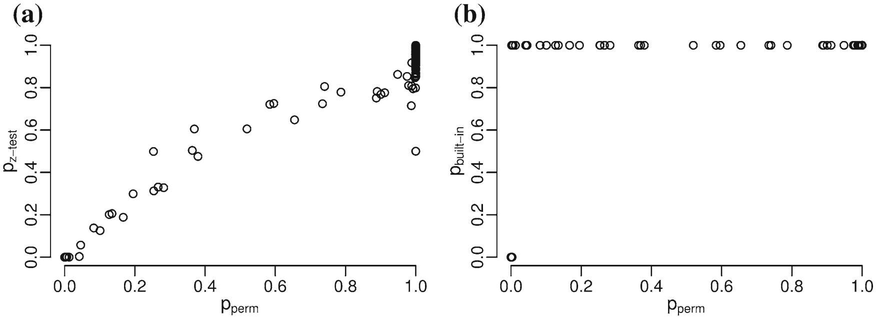


Fig. 4 Agreement between the permutation test and **a** p-values obtained from z-test, or **b** p-values computed using the build-in strategy (showing upper-tail p-values). Calculated for the 100 objects dataset

a way that the influence of the distribution and domain of feature values could be evaluated. We have shown that the RI of a feature depends on its distribution of values across the objects. Features with more evenly distributed values tend to get higher RI scores. This is likely caused by the inclusion of the information gain in the calculation of the RI and may cause trouble if the RIs of all features are assumed to follow the same distribution.

The built-in test in the dmLab software assumes that the RI of all features derive from the same distribution, which may bias the estimate of the feature significances, essentially preventing some features from reaching significance if other—more favorably distributed—features are present in the data. In this study we suggest that each feature should be evaluated individually, using its own null model.

We have shown that a z-test efficiently estimates the correct p-value as validated by a permutation test, whereas applying the built-in strategy combining a t-test with a permutation test failed to detect some significant features and to estimate the "true" p-values obtained by the permutation test. The built-in-strategy treats the RI computed on the original data as a mean instead of a single observation, which may underestimate the sample variation.

It should be noted that since the true standard deviation and mean of the feature RIs on the permuted data is not known, at least 30 permutations have to be performed to convincingly estimate the distribution parameters from the observed data in order to apply a z-test. This puts a lower limit on the number of permutations that can be run to estimate the feature significances. The z-test requires the RIs measured for the permuted data to be approximately normally distributed. Almost all features in our study had a bell shaped distribution, but sometimes with an elongated tail in one direction. Such a tail may lead to an overestimation of the variance in the permuted data, underestimating the significance of a feature. However, we did not observe any such effect.

Since the features are scored according to how they participate in decision tree classifiers, non-informative features will generally not be selected when there are informative features in the same subset. Thus, the more informative features that are present in the data, the lower the non-informative features are scored. We do not expect this effect to significantly affect the estimated p-values of the informative features, but the, comparably, non-informative ones will get very poor p-values, which may explain why many features obtained p-values close to 1 using both the permutation test and the z-test.

Although this methodology is efficient at detecting informative features, the most significant features may not necessarily be the best features to use for classification. The effect size of a feature may be more important than its significance, and both the RI and the p-value should be considered when selecting features for classification.

5 Conclusions

MCFS is a reliable method for feature selection that is able to identify significant features, even with small effects. In this study we showed that features with more evenly distributed values tend to receive higher RIs than features with an uneven distribution. To avoid biasing the selection towards such features, each feature should be tested for significance separately. We have shown that a z-test is an efficient method to estimate the significance of a feature and that these p-values have a strong agreement with p-values obtained through a traditional permutation test.

Acknowledgments We wish to thank the reviewers for insightful comments that helped improve this paper. The authors were in part supported by an ESSENCE grant, by Uppsala University and by the Institute of Computer Science, Polish Academy of Sciences.

References

1. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J. Mach. Learn. Res. 3:1157–1182
2. Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23:2507–2517
3. Draminski M, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J (2008) Monte Carlo feature selection for supervised classification. Bioinformatics 24:110–117
4. Kierczak M, Ginalski K, Draminski M, Koronacki J, Rudnicki W, Komorowski J (2009) A rough set-based model of HIV-1 reverse transcriptase resistome. Bioinform. Biol. Insights 3:109–127
5. Draminski M, Kierczak M, Koronacki J, Komorowski J (2010) Monte Carlo feature selection and interdependency discovery in supervised classification. Stud Comput Intell 263:371–385
6. Enroth S, Bornelöv S, Wadelius C, Komorowski J (2012) Combinations of histone modifications mark exon inclusion levels. PLoS ONE 7:e29911
7. Bornelöv S, Sääf A, Melen E, Bergström A, Moghadam BT, Pulkkinen V, Acevedo N, Pietras CO, Ege M, Braun-Fahrlander C, Riedler J, Doekes G, Kabesch M, van Hage M, Kere J, Scheynius A, Söderhäll C, Pershagen G, Komorowski J (2013) Rule-based models of the interplay between genetic and environmental factors in Childhood Allergy. PLoS ONE 8(11):e80080
8. Kruczyk M, Zetterberg H, Hansson O, Rolstad S, Minthon L, Wallin A, Blennow K, Komorowski J, Andersson M (2012) Monte Carlo feature selection and rule-based models to predict Alzheimer's disease in mild cognitive impairment. J Neural Transm 119:821–831
9. http://www.ipipan.eu/staff/m.draminski/files/dmLab185.zip
10. Van AHT, Saeys Y, Wehenkel L, Geurts P (2012) Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. Bioinformatics 28:1766–1774
11. Dramiński M, Kierczak M, Nowak-Brzezińska A, Koronacki J, Komorowski J (2011) The Monte Carlo feature selection and interdependency discovery is unbiased, vol 40, pp 199–211. Systems Research Institute, Polish Academy of Sciences

ADX Algorithm for Supervised Classification

Michał Dramiński

Abstract In this paper, a final version of the rule based classifier (ADX) is presented. ADX is an algorithm for inductive learning and for later classification of objects. As is typical for rule systems, knowledge representation is easy to understand by a human. The advantage of ADX algorithm is that rules are not too complicated and for most real datasets learning time increases linearly with the size of a dataset. The novel elements in this work are the following: a new method for selection of the final ruleset in ADX and the classification mechanism. The algorithm's performance is illustrated by a series of experiments performed on a suitably designed set of artificial data.

1 Introduction

At present we have a lot of different classification methods. The most popular methods are those based on distance or dissimilarity measure (e.g. kNN [6, 10]), hierarchical methods (e.g. CART [3] and C4.5 [18, 19]), probabilistic methods (e.g. Bayessian classifier), logic systems based on rough sets or classification rules (LEM2, MLEM2 [12, 17, 20], AQ [14, 15], CN2 [4]), and neural nets. Prediction quality of each classifier depends on a problem. After good preparation of input data (feature extraction and selection, discretization if needed, events selection) and optimization of classifier parameters, different classification techniques often give comparable results. For very large datasets a 'perfect' classifier should not need separate feature selection process and is fast enough (learning time is a logarithmic or linear function of size of the dataset). Additionally, a classifier should generalize knowledge after the learning process. The ADX algorithm was designed to meet all these criteria and hopefully, it is not just one more classifier but also a good candidate to deal with very large datasets.

M. Dramiński (✉)
Institute of Computer Science, Polish Academy of Sciences,
Jana Kazimierza 5, 01-248 Warsaw, Poland
e-mail: mdramins@ipipan.waw.pl

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_3

2 Definitions

Let D denote a database. Each row in a database is an event/object e, what means that D is a set of events. Let $|D|$ denote the number of events in D. Each column in database D corresponds to one attribute of events. Attribute can be nominal such as color (e.g., possible values of color are: green, red, etc.) or ordered such as height represented by an ordered set of values/levels (e.g., small, medium and high), or numerical such as height and weight measured, respectively, in inches and pounds. Attributes describe events in D. All possible combinations of values of attributes describing an event form a domain. We assume that D also includes a special attribute called decision attribute **d** which determines class of each event. Thus, class d_i is a value of nominal attribute **d** from a finite set $d_i = \{d_1, \ldots, d_m\}, m < \infty$.

Selector s is an expression that describes set of events by imposing some condition on values one attribute. For example: $color = blue$; or $weight > 80$. Each simple selector consists of a name of attribute, its single value and an operator relating the two. Advanced/complex selector can include a list of nominal or ordered values (e.g. $color = [blue, red]$) or a range of numeric values (e.g. $weight = (70; 80]$). Each complex selector can be written as a set of simple selectors suitably combined into one selector: $color = [blue, red]$ is equivalent to $color = blue$ OR $color = red$, $weight = (70; 80]$ is equivalent to $weight > 70$ AND $weight \leq 80$. Selector s which allows any value for the attribute is called universal/general and will be presented as $attribute = *$ (e.g. $color = *$ denotes any color).

Let complex c denote a set of selectors and let length n of the complex denote the number of selectors (simple or advanced) contained in the complex. For example complex $c = \langle s_1, s_2, s_3 \rangle$, has length 3 (where $s_j, j = \{1, 2, 3\}$, is a selector). A complex is understood as a conjunction of selectors.

Coverage of selector s, $cov(s)$ is the number $|D^s|$ of events that satisfy condition in s divided by $|D|$. Coverage of c, $cov(c)$ is the number $|D^c|$ of events that satisfy complex c divided by $|D|$.

$$cov(s) = \frac{|D^s|}{|D|}, cov(c) = \frac{|D^c|}{|D|} \tag{1}$$

For a given class d_i, positive set $D_{d_i,p}$ is the set of events whose decision attribute has the value corresponding to this class. Accordingly, during creation of rules, for a given class d_i, all events that belong to d_i are called positive, all other events are called negative $D_{d_i,n}$. For a given class, positive coverage $pcov_{d_i}$ of a given complex c is the coverage on the subset of events which belong to the considered class d_i. Negative coverage $ncov_{d_i}$ of the complex c is the coverage on the subset of events whose class is different from the considered class d_i, thus

$$pcov_{d_i}(c) = \frac{|D^c_{d_i,p}|}{|D_{d_i,p}|}, ncov_{d_i}(c) = \frac{|D^c_{d_i,n}|}{|D_{d_i,n}|} \tag{2}$$

$$pcov_{d_i}(s) = \frac{|D^s_{d_i,p}|}{|D_{d_i,p}|}, ncov_{d_i}(s) = \frac{|D^s_{d_i,n}|}{|D_{d_i,n}|} \tag{3}$$

where $D^c_{d_i,p}$ denotes the set of positive events for class d_i, covered by complex c. Analogically, $D^c_{d_i,n}$ denotes the set of negative events for class d_i, covered by c, and $D^s_{d_i,p}$, $D^s_{d_i,n}$ have the suitable meaning for selector s. Clearly $D^c_{d_i,p} \cup D^c_{d_i,n} = D^c$. Note that for any d_i:

$$cov(s) = \frac{|D^s_{d_i,p}| + |D^s_{d_i,n}|}{|D|}, cov(c) = \frac{|D^c_{d_i,p}| + |D^c_{d_i,n}|}{|D|}, \tag{4}$$

The probability of positive class occurrence for complex c is given by Eq. (5).

$$prob_{d_i,p}(c) = \frac{|D^c_{d_i,p}|}{|D^c_{d_i,p}| + |D^c_{d_i,n}|} \tag{5}$$

By definition, strong rules characterize classes uniquely, and hence their negative coverage $ncov$ is equal to 0. If a complex has $ncov > 0$ and $pcov > ncov$ for a given class d_i it still characterizes this class but with some probability lesser than 1. This probability is usually called confidence or consistency of a rule. Confidence for any strong rule equals 1. A set of complexes, understood as disjunction of the complexes, that characterizes one class with positive confidence is called a ruleset. A set of rulesets that characterizes all classes with positive levels of confidence is called a ruleset family.

3 The ADX Algorithm—Creation of a Ruleset for One Class

Let us begin with a loose description of the algorithm. The idea of the ADX (**A**priori–**D**ecision rules e**X**traction) algorithm is based on the observation that conjunction of selectors cannot have a larger cov than minimal cov for each of these selectors. The main goal of the algorithm is to find the best ruleset (the set of complexes such that each of them has a possibly large $pcov$ and a possibly small $ncov$ for each of the classes). The rules are built by an iteration process. In each iteration, we lengthen complexes of high quality by one selector. Quality measure is based on $pcov$ and $ncov$ and roughly speaking it is high if $pcov$ is high and $ncov$ is low. After creating all the rulesets, classification can be done by measuring similarity of a new unlabeled event to each ruleset. The ruleset which is most similar to the considered event will be chosen as a classifier's decision. We allow ambiguities, noise, and missing values in input data.

The ADX algorithm does not try to create very long and complicated rules with $ncov = 0$ (strong rules). The idea is to create a set of rather simple rules with possibly large $pcov$ and possibly small $ncov$. Sometimes a set of strong rules can perform badly during later prediction (e.g. for noisy data), and therefore ADX does not treat strong rules in any special way. In the algorithm, there is no need to truncate the rules finally obtained since the way they are constructed makes them possibly simple, short and general. We assume in the sequel that input data are discretized beforehand, and hence that the ADX works on nominal values. However, a very good discretization technique as proposed by Fayyad and Irani [9] is implemented into ADX.

3.1 First Step—Finding Selectors Base

For a given class and for each value of each attribute in D, the algorithm finds $pcov$ and $ncov$. It means that in the first step, the algorithm builds a set of contingency tables for all attributes (excluding $\mathbf{d}$). The set of all simple selectors thus obtained, to be termed selectors base, can be sorted by quality measure Q which combines $pcov$ and $ncov$ (definition of Q will be given later).

3.2 Second Step—Creation of Complexes Set

We can say that the selectors base is a set of complexes with length 1. Longer complexes are created via an iteration process. Each iteration consists of:

- Creation of candidates (each of them being 1 simple selector longer than selectors in the previous iteration).
- Estimation of candidates' quality.
- Deleting useless complexes.
- Selection of parents set.
- Checking if stop criteria have been reached.

Creation of Candidates Complexes whose quality has not been estimated yet are called candidates. Creation of such complexes with length 2 is very simple. These are all possible selectors' pairs excluding pairs where both selectors are based on the same attribute. Notice that complex (length 2) $c_1 = \langle s_1, s_2 \rangle$ equals $c_2 = \langle s_2, s_1 \rangle$, because each complex consists of conjunction of selectors. There is an important issue about creation of complexes longer than 2. To create a new complex candidate of length $n + 1$, the algorithm uses two complexes shorter by 1, which can be called parents. Parent complexes need to have common part of length $n - 1$ (where n denotes length of parents complexes). For example complex $c_1 = \langle s_1, \mathbf{s_2}, \mathbf{s_3} \rangle$ and $c_2 = \langle \mathbf{s_2}, \mathbf{s_3}, s_5 \rangle$ can be used to create $c_3 = \langle s_1, \mathbf{s_2}, \mathbf{s_3}, s_5 \rangle$. Creation of complexes with length 2 is based on simple selectors. Complexes with length 3 are based on selected parents complexes with length 2 and so on.

Evaluation of Candidates' Quality After creation of the candidates set there is a need to separately evaluate quality of each newly created complex. During the evaluation process, for each complex candidate there is calculated positive and negative coverage ($pcov$ and $ncov$) on the training set. Based on these coverages, the quality measure Q_1 is calculated and is used to evaluate complexes' usability.

$$Q_1 = (pcov - ncov)(1 - ncov) \tag{6}$$

Deleting Useless Candidates Complexes that do not cover any positive event ($pcov = 0$) are deleted (to save the memory) during this step. Such complexes do not describe a class and are useless for next steps. However the rest of the complexes, including those that will not be used for creation of new candidates, are stored for future final selection step.

Selection of Parents Within this step, ADX selects complexes as parents from which, next candidates (longer by 1) are created. The size of parents set is defined by parameter $searchBeam$ and selection of parents is based on measure Q_1 (higher is better). If the number of complexes which have the same (and the highest) value of Q_1 is larger than $searchBeam$ then exactly $searchBeam$ parents are selected at random. Complexes which were not selected to be parents in the next iteration, are stored to be used for final selection process. While parameter $searchBeam$ controls the scope of exploration, it affects the learning time. With $searchBeam$ increasing the scope of exploration increases, but unfortunately, learning time grows too. Loosely speaking, $searchBeam$ should be set in such a way that possibly best results be obtained within acceptable time. Its default value is set to 50.

In this phase, selection of complexes with $ncov = 0$ as parents is allowed. Experiments have shown that it leads to building longer complexes which can play positive role during the classification process (the ruleset proves more specific).

Stop Criteria Creation of candidates has two possible stop criteria. According to the first criterion, creation of new candidates is stopped when complex length equals the number of attributes (without decision attribute). Creation of the ruleset has to be finished because each single selector in complex must describe a different attribute (complex cannot be any longer).

According to the second criterion, the algorithm is stopped when the set of new candidates is empty (if there are no parents that have common part of $n - 1$ length and new candidates cannot be created).

3.3 Third Step—Merging of Complexes

In this step, the number of created and evaluated complexes is decreased and their quality improved. If some complexes are based on the same attribute set and only one selector has different value, then it is possible to merge such complexes. Merging induces creation of new selectors and removal of those merged. This new selector is a disjunction of values of the corresponding base selectors. For example: $\langle A = 1, B = 3, C = 4\rangle \oplus \langle A = 1, B = 3, C = 7\rangle \Longrightarrow \langle A = 1, B = 3, C = [4, 7]\rangle$, where new selector is $C = [4, 7]$. For the resulting complex $pcov$ and $ncov$ are the sums of corresponding coverages of removed complex.

3.4 Fourth Step—Final Selection of the Ruleset

The final selection step is critical for the ruleset creation and is similar to truncation of a tree—essentially it has the same effect. From the entire set of stored rules we have to select the most appropriate subset that can be used in later successful classification.

In this phase, from the entire set of complexes, maximally $finalBeam$ (a number not larger than fixed value of this parameter) of complexes is selected. Selection is based on prediction ability of already selected rules. First of all, all the complexes discovered so far are sorted by quality measure Q_2 (7). The measure Q_2 is similar to Q_1 but is more focused on complexes with lower $ncov$.

$$Q_2 = (pcov - ncov)(1 - ncov)^2 \tag{7}$$

In order to describe final selection process, a measure Q_r of prediction ability of a ruleset has to be introduced.

$$Q_r = (S^p - S^n)(1 - S^n) \tag{8}$$

where

$$S^p = \sum_{e \in D_p} S(e) \tag{9}$$

$$S^n = \sum_{e \in D_n} S(e) \tag{10}$$

and $S(e)$ is measure of event e, to be defined in the next section.

The idea of measure Q_r is to estimate prediction ability of a ruleset on a given subset of events. Factor S^p denotes the sum of scores over all positive events in the subset. Loosely speaking score $S(e)$ of the event e expresses the level of belonging to the given ruleset. Analogously, factor S^n is calculated only for negative events from the selected training subset.

After sorting all the complexes we can start from the complex with the highest Q_2, and add complexes, one by one, to the temporary final ruleset. If measure Q_r calculated for the entire temporary ruleset does not decrease, the temporarily added rule joins the final ruleset. If not the rule is removed from the temporary ruleset and the next rule from the ranking (with slightly lower Q_2) is temporarily added. If the training set is very large to speed up the final selection process we can estimate Q_r

on a randomly selected subset of training events. The size of the selected subset is defined by parameter $maxEventsForSelection$.

Measure Q_r has high value if rules contained in temporary ruleset have possibly high score for all positive events in the subset data and possibly low score for negative events in the same subset. Generally speaking, the selection of rules to the final ruleset is based on optimization of successful reclassification of events from the training subset.

Notice that the number and quality of selected rules have high impact on speed and quality of later classification. Therefore maximal number of rules in the final ruleset (parameter $finalBeam$) can be set larger than its default value for extremely large and complicated data. However, for some datasets, if Q_r cannot be increased by adding next complexes, the size of the final ruleset can be smaller than default $finalBeam$ value (e.g. if 1 rule in a given ruleset is sufficient to successfully predict the class). The default setting for $finalBeam$ is 50.

4 Classification of New Events

During the classification process the system calculates the set of scores $S_{d_i}(e)$ which describes strength of evidence of the considered event e belonging to class d_i.

1.

$$S_{d_i}(e) = \begin{cases} S'_{d_i}(e) = \frac{\sum_j prob_{d_i,p}(r^j_{d_i}(e))}{|r_{d_i}(e)|} \\ \frac{p_{d_i}(e)}{p_{d_i}} \text{ if } \exists_{S'_{d_1}(e),S'_{d_2}(e)} S'_{d_1}(e) = S'_{d_2}(e) = max(S'_{d_i}(e)) \end{cases}$$

where:

- $r_{d_i}(e)$—denotes the set of rules/complexes from class d_i that cover an event e
- $p_{d_i}(e)$—denotes the sum of $pcov_{d_i}$ of rules that cover an event e
- p_{d_i}—denotes the sum of $pcov_{d_i}$ of rules from class d_i
- $prob_{d_i,p}(r^j_{d_i}(e))$—denotes probability of class occurrence under condition that rule/complex $r^j_{d_i}$ covers the event e

For each class, the system calculates separate score value. Class d_i that has maximal value of a calculated index $S_{d_i}(e)$ provides the final decision. Score $S'_{d_i}(e)$ represents average probability of rules (from one ruleset) that cover the event. If maximal value of $S'_{d_i}(e)$ is related to more than one class (e.g. if the event is covered by only strong rules $S'_{d_1}(e) = S'_{d_2}(e) = 1$), the score $S_{d_i}(e)$ is calculated based on suitable sums of $pcov_{d_i}$. In this case the score is equal to the fraction of the sum of $pcov_{d_i}(e)$ of rules that cover the event and the sum of $pcov_{d_i}$ of all rules in the ruleset created for class d_i. If $S_{d_i}(e)$ still gives no result (e.g. the score for d_1 equals to score for d_2) then the event is randomly classified based on frequency distribution of considered classes. However this situation in extremely rare.

Factors $pcov_{d_i}(c)$ and $prob_{d_i,p}(c)$ are calculated and stored for all complexes from the final set. If an event is covered by a rule, the system uses suitable factor that have been found during the learning process and based on that $S_{d_i}(e)$ and $S'_{d_i}(e)$ are calculated.

In previous papers [7, 8] some different indices have been proposed but after series of experiments $S_{d_i}(e)$ has proven to be most stable in giving best prediction results and the ADX algorithm uses $S_{d_i}(e)$ (as a default) to calculate score.

5 Experiments

5.1 Artificial Data

In order to check prediction ability of the proposed classifier special synthetic data were prepared. Based on **R** environment [21] 4 data sets were prepared $k = 800, 4000, 8000, 16000$ and each data set contained k events split into two equal size classes.

The data contained 3 attributes: 2 numerical(x, y) that describe location in 2D space and the third one—decision attribute that determines the class: *blue*, *red*. For the class *blue* values of x and y were generated with normal distribution ($mean = 0.5$ and $stdev = 0.1$). The class *red* is based on 4 similar normal distributions (each of them contained $\frac{n}{8}$ of events) with $stdev = 0.1$ and placed in the space in such a way as to surround class *blue* (see Fig. 1).

The ruleset built on all events of the dataset with $k = 800$ is the following:

```
#Rules for decision = blue
 #Complex Size: 2
  x=(0.25748622;0.35888457] and y=(-Infinity;0.7059473]p:0.087
    n: 0.002 q: 0.0847 pr: 0.9722
```

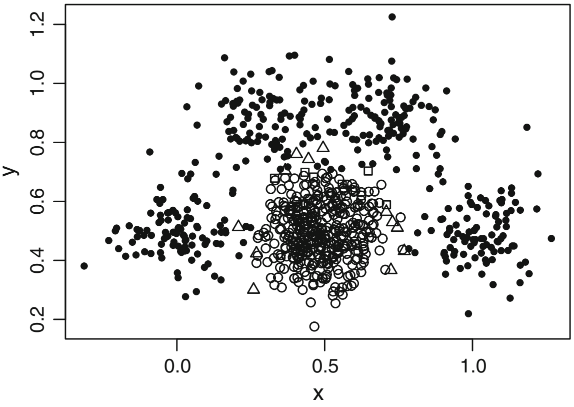


Fig. 1 Result of ADX cross validation, dataset $n = 800$, ●—correctly classified *red*, ○—correctly classified *blue*, △—incorrectly classified *blue*, □—incorrectly classified *red*

```
  x=(0.35888457;0.6413076] and y=(-Infinity;0.7059473] p: 0.832
    n: 0.012 q: 0.8097 pr: 0.9852
  x=(0.6413076;0.7087706] and y=(-Infinity;0.7059473] p: 0.054
    n: 0.002 q: 0.0523 pr: 0.9565
  x=(0.7087706;0.7711703] and y=(-Infinity;0.7059473] p: 0.014
    n: 0.002 q: 0.0124 pr: 0.8571

#Rules for decision = red
 #Complex Size: 1
  x=(-Infinity;0.25748622]p: 0.340 n: 0.002 q: 0.3366 pr:0.9927
  x=(0.7711703;Infinity] p: 0.302 n: 0.0 q: 0.3025 pr: 1.0
  y=(0.7059473;Infinity] p: 0.492 n: 0.007 q: 0.4813 pr: 0.9850
 #Complex Size: 2
  x=(0.25748622;0.35888457] and y=(0.7059473;Infinity] p: 0.072
    n: 0.0 q: 0.0724 pr: 1.0
  x=(0.6413076;0.7087706] and y=(0.7059473;Infinity] p: 0.052
    n: 0.0 q: 0.0524 pr: 1.0
  x=(0.7087706;0.7711703] and y=(0.7059473;Infinity] p: 0.072
    n: 0.0 q: 0.0724 pr: 1.0
  x=(0.7711703;Infinity] and y=(-Infinity;0.7059473] p: 0.242
    n: 0.0 q: 0.2425 pr: 1.0
```

Values: p, n denote $pcov$ and $ncov$, q quality Q_1 of the rule and pr probability of positive class occurrence.

Results of 3 fold cross validation (average accuracy calculated for 10 repetitions) processed on 4 datasets of different sizes are summarized in Table 1. The ADX classifier was implemented in JAVA 1.6 and experiments were run on standard PC (Pentium IV 3.2 GHz, 4 GB RAM) running WinXP. All other algorithms came from Weka [22] ver. 3.4.11 and were run under **R** (library RWeka) with their default settings. The SVM(SMO) classifier was run under slightly changed parameters ($C = 5.0$,

Table 1 Average accuracy and processing time of a single 3-fold *cv*

k	J48	1-NN	NaiveBayes	RandomForest	SVM	**ADX**
800	97.9	96.8	96.7	97.6	94.8	**97.4**
Time[s]	0.19	0.37	0.14	0.51	0.74	**0.08**
4000	97.6	97.0	96.0	97.3	97.3	**97.3**
Time[s]	0.62	3.08	0.31	2.14	22.38	**0.44**
8000	97.7	97.0	96.6	97.2	97.4	**98.0**
Time[s]	1.3	12.3	0.5	5.1	91.95	**1.16**
16000	97.6	96.7	96.3	NaN	97.4	**97.8**
Time[s]	2.9	56.9	1.4	32.3	438.9	**3.74**

Table 2 Average accuracy and processing time of a single 3-fold *cv* (added noisy attributes)

k	J48	1-NN	NaiveBayes	RandomForest	SVM	**ADX**
800	97.4	56.8	95.7	94.3	67.5	**97.3**
Time[s]	0.9	1.9	0.6	2.4	23	**0.36**
4000	97.1	54.2	95.9	97.0	89.6	**97.3**
Time[s]	5.9	33.3	3.2	16.2	4314	**4.05**
8000	97.1	56.7	96.5	97.4	95.4	**98.0**
Time[s]	13.7	122.2	6.4	36.9	11015	**13.69**
16000	97.1	55.8	96.3	97.4	96.8	**97.8**
Time[s]	30.2	438.2	12.5	82.4	29046	**50.21**

$exponent = 2.0$, $lowerOrderTerms = true$) because for the specified classification problem linear polynomial kernel has to be used.

Prediction accuracies for almost all algorithms are very high and also similar to each other. Time needed to process 3 fold cross validation is more diversified and ADX compared to others techniques seems to be very effective. Figure 1 presents results of one cross validation realization (data set $k = 800$).

For the second set of experiments each of previous datasets contained additionally 30 new randomly generated attributes (each with uniform distribution from the range [0; 1]). The intention of such addition was to examine if ADX is sensitive to non informative attributes and what kind of impact these attributes can have on created rules. Table 2 presents results of experiments that were processed in the same way as previously. The SVM weka implementation is not efficient when it deals with many non informative attributes and it can be easily noticed comparing presented cv times.

Noisy attributes badly affected only on the performance 1-NN classifier. It is no surprise because k-NN uses all attributes to measure the distance and without sensible feature selection k-NN leads to low prediction quality. For the ADX classifier results did not change and rules are based only on informative x and y attributes. For the given set of experiments the ADX was also one of the fastest classifiers.

5.2 *Real Data*

The second set of experiments has performed on real data, most often from [16]. For the experiments, the author used also a few commercial datasets that cannot be published, but the results of the classification can show differences between classifiers (Table 3).

The dataset ‘frauds’ contains fraudulent transactions of reward receiving in loyalty program VITAY of PKN ORLEN (PKN ORLEN is the biggest oil company in central Europe). The number of fraudulent transactions is 3072. The dataset ‘Clients’ contains customer profiles that are divided into 6 classes. The smallest class contains

Table 3 Real datasets used for classification experiment

Dataset	Events	Attributes	Classes	Source
Alizadehdata	62	4026	3	[2]
Golubdata	38	7129	2	[11]
Frauds	15123	33	2	PKN Orlen
Mushroom	8123	22	2	[16]
Iris	150	4	3	[16]
AbaloneData	4177	8	3	[16]
Soybean	683	35	19	[16]
Tic_tac_toe	958	9	2	[16]
Hypothyroid	3772	30	4	[16]
Ionosphere	351	35	2	[16]
Splice.asx	3190	61	3	[16]
Vote	435	17	2	[16]
Zoo	101	18	7	[16]
Clients	16920	59	6	real Polska

919 profiles and the largest contains 7598 profiles. Each customer is described by a set of demographic features (e.g. sex, age etc.) as well as by behavioral features based on transaction history (e.g. mobility, loyalty etc.).

For each of the datasets 3-fold cv was used 10 times of. The average result of 10 cross validations is given in Table 4. Classifiers: NaiveBayes (NB), C4.5 (J48), SVM(SMO), kNN (IBk) implemented in WEKA ver. 3.4.11, were run on their default parameters (kNN: 1-NN, linear kernel in SVM). Implementation of ADX, done in JAVA 1.6, was run on its default parameters too. The factor wAcc is the well known weighted/balanced accuracy which is sensitive to problems where classes are highly unbalanced. Time given in the table is the average time of a single 3-fold cross validation process.

Classification quality (see Table 4) of the ADX implementation is comparable to that achieved by popular effective classifiers and for the most datasets the ADX is one of the best classifiers. However there are a few datasets where default parameters of ADX did not lead to the highest performance. However, by applying slightly different method for final selection of rules for ‘tic_tac_toe’ we can achieve average wAcc = 0.985, for ‘Soybean’ average wAcc = 0.864 and for ‘zoo’ average wAcc = 0.811.

The time needed for processing a single cross validation proces is much more diversified. Great scalability of ADX can be noticed especially for large datasets. For the dataset ‘frauds’ and ‘Clients’ average cv time equals to several/tens of seconds in comparison to hundreds/thousands of seconds for other classifiers. In the case of smaller datasets the ADX algorithm is still one of the fastest techniques but the difference is not such significant.

Table 4 Results of classification of real datasets

Dataset	ADX	J48	1-NN	SVM	NB
Alizadeh Acc(wAcc)	94.8(89)	83.7(72.2)	97.7(95.5)	99.8(99.6)	89.4(77)
Alizadeh CV[s]	0.77	1.44	1.68	1.15	1.21
Golub Acc(wAcc)	87.4(79.3)	83.9(82)	87.9(81.5)	92.9(87.7)	92.6(87.3)
Golub CV[s]	0.7	1.2	1.2	1.5	1
frauds Acc(wAcc)	85.7(74.4)	86.7(76.8)	83.5(76.2)	86.3(73.7)	47.9(64.3)
frauds CV[s]	6.4	24.6	831.9	196.5	2.9
mushroom Acc(wAcc)	98.6(98.5)	100(100)	100(100)	100(100)	95.4(95.3)
mushroom CV[s]	2.3	0.3	139.1	19	0.2
iris Acc(wAcc)	93.2(93.2)	94.1(94.1)	95(95)	96.4(96.4)	95.5(95.5)
iris CV[s]	0.01	0.01	0.02	1.93	0.01
abalone Acc(wAcc)	52.4(52)	52.7(52.7)	50.1(50.2)	54.5(53.1)	51.9(53.7)
abalone CV[s]	1.64	1.46	16.42	1.24	0.25
soybeam Acc(wAcc)	81.9(86.4)	88.9(86.5)	90.9(91.9)	92.5(95.3)	90.6(91.8)
soybeam CV[s]	2.3	0.1	1.4	121	1.4
tic_tac Acc(wAcc)	76.9(71)	83.3(79.7)	97.5(96.4)	98.3(97.6)	70.9(64)
tic_tac CV[s]	0.16	0.05	0.85	1.22	0.02
hypothyroid Acc(wAcc)	97.7(67.3)	99.5(72.4)	91.5(43.7)	93.6(38.3)	95.3(53.3)
hypothyroid CV[s]	2.2	0.4	38.8	17.4	0.3
ionosphere Acc(wAcc)	87.3(87.8)	89.7(87.7)	86.4(82)	87.9(84.2)	82.7(83.5)
ionosphere CV[s]	0.2	0.25	0.47	0.32	0.06
splice.asx Acc(wAcc)	86.1(85.7)	93.4(93.4)	73.4(78.9)	73.4(79)	95.3(94.8)
splice.asx CV[s]	1.9	0.5	51.8	54.4	0.2
vote Acc(wAcc)	94.1(94.2)	95.7(95.5)	92.2(92.6)	95.7(95.8)	90.1(90.4)
vote CV[s]	0.2	0.03	0.31	0.13	0.01

(continued)

Table 4 (continued)

Dataset	ADX	J48	1-NN	SVM	NB
zoo Acc(wAcc)	83.8(72.8)	93.4(85.5)	95(89.9)	95.2(88.1)	95(91)
zoo CV[s]	0.16	0.01	0.06	1.43	0.01
Clients Acc(wAcc)	81.6(73.6)	89.6(85.5)	76.2(68.6)	87(83)	76.1(68.5)
Clients CV[s]	31	50	1739	483	1759

6 Conclusions

In the process of optimization of classifier's parameters the speed of learning and testing is very important. Today, in practice, we deal with huge amount of data and even the data sample often contains thousands of events. Therefore, if various techniques have comparable classification quality, the scalability of classifier in many practical applications is the most important. The ADX algorithm gives comparable results of classification to the well known popular classification techniques but its scalability is much better.

The general idea of combining single selectors to lengthen the rules in ADX is very similar to lengthen itemsets in Apriori algorithm [1], but ADX produces classification rules and uses combination of *pcov* and *ncov* to estimate rules quality. The quality Q of the rules is used to select a fixed number of best complexes (parents) to build complexes candidates. This solution has positive influence on efficiency of algorithms because it limits the number of possible candidates to be created. It also helps to create rules that do not need post pruning because high quality complex still can cover some of negative events. The main idea of specialization of rules to increase their quality is commonly used in many algorithms that are based on sequential covering schema (e.g. AQ, LEM2, CN2, RIPPER [5] etc.). Creation of many rules by combining simple selectors is in some sense analogous to creation of spline functions in MARS (Multivariate Adaptive Regression Splines Model [13]) by combining basis functions. However, the ADX algorithm is a new rule classifier algorithm that is fast and highly efficient, what makes it very attractive considering large datasets.

References

1. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th international conference on very large databases, Santiago, Chile
2. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudso J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R,

Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) Distinct types of diffuse large B-cell Lymphoma identified by expression profiling. Nature 403:503–511

3. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International Group, Monterey
4. Clarc P, Niblett T (1989) The CN2 induction algorithm. Mach. Learn. 3:261–283
5. Cohen W (1995) Fast effective rule induction. In: Machine learning: proceedings of the twelfth international conference, Lake Tahoe, California
6. Cover TM, Hart PE (1967) Nearest neighbor pattern classification. IEEE Trans. Inform. Theory, IT-13(1):2127
7. Dramiński M (2004) ADX Algorithm: a brief description of a rule based classifier. In: Proceedings of the new trends in intelligent information processing and web mining IIS'2004 symposium. Springer, Zakopane, Poland
8. Dramiński M (2005) Description and practical application of rule based classifier ADX, proceedings of the first Warsaw International Seminar on intelligent systems, WISIS (2004). In: Dramiski M, Grzegorzewski P, Trojanowski K, Zadrozny S (eds) Issues in intelligent systems. Models and techniques. ISBN 83-87674-91-5, Exit 2005
9. Fayyad UM, Irani KB (1992) On the handling of continuous-valued attributes in decision tree generation. Mach. Learn. 8:87–102
10. Fix E, Hodges JL (1951) Discriminatory analysis nonparametric discrimination: Consistency properties, Project 21–49-004, Report no. 4, USAF School of Aviation Medicine, Randolph Field, pp 261–279
11. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh M, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of Cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537
12. Grzymala-Busse JW (2003) MLEM2-Discretization during rule induction, intelligent information processing and web mining. In: Proceedings of the international IIS:IIPWM'03 conference held in Zakopane, Poland
13. Hastie T, Tibshirani R, Friedman J (2001) Elements of statistical learning: data mining, inference and prediction. Springer, New York
14. Kaufman KA, Michalski RS (1999) Learning from inconsistent and noisy data: the AQ18 approach. In: Proceedings of the eleventh international symposium on methodologies for intelligent systems (ISMIS'99), Warsaw, pp 411–419
15. Michalski RS, Kaufman KA (2001) The AQ19 system for machine learning and pattern discovery: a general description and user's guide, reports of the machine learning and inference laboratory, MLI 01–2. George Mason University, Fairfax
16. Newman DJ, Hettich S, Blake CL, Merz CJ (1998) UCI Repository of machine learning databases. University of California, Department of Information and Computer Science, Irvine, http://www.ics.uci.edu/mlearn/MLRepository.html
17. Pawlak Z (1991) Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publishing, Dordrecht
18. Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann
19. Quinlan JR (1986) Induction of decision trees. Machine learning 1:81–106
20. Stefanowski J (1998) On rough set based approaches to induction of decision rules. In: Polkowski L, Skowron A (eds) Rough sets in data mining and knowledge discovery, Physica-Verlag, pp 500–529
21. The R project for statistical computing. http://www.r-project.org/
22. Wittenn IH, Eibe F (2005) Weka 3: data mining software in Java, data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco

Estimation of Entropy from Subword Complexity

Łukasz Dębowski

Abstract Subword complexity is a function that describes how many different substrings of a given length are contained in a given string. In this paper, two estimators of block entropy are proposed, based on the profile of subword complexity. The first estimator works well only for IID processes with uniform probabilities. The second estimator provides a lower bound of block entropy for any strictly stationary process with the distributions of blocks skewed towards less probable values. Using this estimator, some estimates of block entropy for natural language are obtained, confirming earlier hypotheses.

Keywords Subword complexity · Block entropy · IID processes · Natural language · Large number of rare events

1 Introduction

The present paper concerns estimation of block entropy of a stationary process from subword complexity of a sample drawn from this process. The basic concepts are as follows. Fix $\mathbb{X}$ as a finite set of characters, called alphabet. Let $(X_i)_{i\in\mathbb{Z}}$ be a (strictly) stationary process on a probability space $(\Omega, \mathcal{J}, P)$, where $X_i : \Omega \to \mathbb{X}$ and the blocks are denoted $X_k^l = (X_i)_{i=k}^l$ with probabilities $P(w) := P(X_1^{|w|} = w)$. Function $H(k) := \mathbf{E}\left[-\log P(X_1^k)\right]$ is called block entropy. It is nonnegative, growing, and concave [3, 4]. Let λ denote the empty string. Elements of set $\mathbb{X}^* = \{\lambda\} \cup \bigcup_{n\in\mathbb{N}} \mathbb{X}^n$ are called finite strings. For a given string $w \in \mathbb{X}^*$, substrings of w are finite blocks of consecutive characters of w. By $f(k|w)$ we will denote the number of distinct substrings of length k of string w. Function $f(k|w)$ is called subword complexity [13, 19, 25, 28].

Ł. Dębowski (✉)
Institute of Computer Science, Polish Academy of Sciences,
Ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
e-mail: ldebowsk@ipipan.waw.pl

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_4

We are interested how to estimate block entropy $H(k)$ given subword complexity $f(k|X_1^n)$. Estimating block entropy from a finite sample X_1^n is nontrivial since there are $(\operatorname{card} \mathbb{X})^k$ different blocks of length k so we cannot obtain sufficiently good estimates of individual probabilities of these blocks for $(\operatorname{card} \mathbb{X})^k > n$. We expect, however, that this result may be improved for a stationary ergodic process, since by the Shannon-McMillan-Breiman theorem [1], there are roughly only $\exp(H(k))$ different substrings of length k that appear with substantial frequency in realization $(X_i)_{i\in\mathbb{Z}}$. Hence it might be possible to obtain reliable estimates of block entropy for $\exp(H(k)) \le n$. Here we will show that some estimates of this kind could be obtained via subword complexity.

Estimating block entropy for relatively long blocks drawn from a discrete stationary process has not been much investigated by mathematicians. What makes the studied topic difficult is the necessity of doing statistical inference in the domain of large number of rare events (LNRE), cf. [21], and under unknown type of dependence in the process, except for the assumption of strict stationarity. These conditions may preclude usage of standard statistical techniques of improving estimators such as smoothing or aggregation [5, 24]. In the mathematical literature, there are more publications on entropy estimation in the context of entropy rate, e.g., [22, 29, 31, 32], or differential entropy, e.g., [16, 20]. The idea of estimating block entropy for relatively long blocks has been pursued, however, by physicists in some applied works concerning the entropy of natural language and DNA [10–12, 26]. Subword complexity was also used to compute topological entropy of DNA, a somewhat different concept [23].

In fact, the subject of this paper can be motivated by the following applied problem. In the early days of information theory, Shannon [27] made a famous experiment with human subjects and produced estimates of conditional entropy, equal to $H(k+1) - H(k)$, for texts in natural language for block lengths in range $k \in [1, 100]$. Many years later, Hilberg [17] reanalyzed these data and, for the considered k, he found the approximate relationship

$$H(k) \approx Ak^{\beta} + hk \tag{1}$$

with entropy rate $h \approx 0$ and exponent $\beta \approx 1/2$. Moreover, he conjectured that relationship (1) may be extrapolated for much larger k, such as k being the length of a book. There are some rational arguments that such relationship may hold indeed [6] but more experimental support is required. Hence, whereas experiments with human subjects are costly and may be loaded with large errors, there is some need for a purely statistical procedure of estimating block entropy for relatively large blocks.

Approaches to the concerned problem proposed so far were quite heuristic. For example, Ebeling and Pöschel [10] implemented the following scheme. First, let $n(s|w)$ be the number of occurrences of substring s in a string w. For a sample X_1^n, let us consider this naive estimator of entropy,

$$H_{\text{est}}(k) = -\sum_{w\in\mathbb{X}^k} \frac{n(w|X_1^n)}{n-k+1} \log \frac{n(w|X_1^n)}{n-k+1} \ . \tag{2}$$

The estimator is strongly biased. In particular, $H_{\mathrm{est}}(k) \le \log(n-k+1)$. The bias of $H_{\mathrm{est}}(k)$ was corrected in this way. For a Bernoulli process, the expectation of estimator $H_{\mathrm{est}}(k)$ can be approximated as

$$\mathbf{E}\, H_{\mathrm{est}}(k) \approx \begin{cases} H(k) - \frac{1}{2} \frac{\exp H(k)}{n-k+1} \ , & \dfrac{\exp H(k)}{n-k+1} \ll 1 \ , \\ \log(n-k+1) - \log(2) \dfrac{n-k+1}{\exp H(k)} \ , & \dfrac{\exp H(k)}{n-k+1} \gg 1 \ , \end{cases} \tag{3}$$

whereas the value of $\mathbf{E}\, H_{\mathrm{est}}(k)$ for k between these two regimes can be approximated by a Padé approximant. Hence, given an observed value of $H_{\mathrm{est}}(k)$ and assuming that it is equal $\mathbf{E}\, H_{\mathrm{est}}(k)$ and falls between the two regimes, $H(k)$ was estimated by inverting the Padé approximant. The estimates obtained in [10] suggest that block entropy for texts in natural language satisfies Hilberg's hypothesis (1) for $k \le 25$. One may doubt, however, whether the obtained estimates can be trusted. First, natural language is not a Bernoulli process and, second, using the Padé approximant instead of a rigorously derived expression introduces unknown errors, which can explode when inverting the approximant.

Consequently, in this paper we will pursue some new mathematically rigorous ideas of estimating block entropy from subword complexity. We propose two simple estimators of block entropy. The first estimator works well only in the simple case of IID processes with uniform probabilities. Thus we propose a second estimator, which works for any stationary process for which the distribution of strings of a given length is asymmetric and skewed towards less probable values. It should be noted that this estimator yields a lower bound of entropy, in contrast to estimators based on universal source coding, such as the Lempel-Ziv code, which provide an upper bound of entropy [7, 9, 31, 32]. Using the second estimator, we also estimate block entropy for texts in natural language, confirming Hilberg's hypothesis (1) for $k \le 10$. We believe that this result might be substantially improved. We suppose that subword complexity conveys enough information about block entropy for block lengths smaller than or equal the maximal repetition. For natural language, this would allow to estimate entropy for $k \le 100$ [8].

2 Theoretical Results

Our starting point will be formulae for average subword complexity of strings drawn from some stochastic processes. A few such formulae have been given in [14, 18, 19]. We will derive a weaker but a more general bound. First of all, let us recall, cf. [25], that function $f(k|X_1^n)$ for a fixed sample X_1^n is unimodal and k for which $f(k|X_1^n)$ attains its maximum is called the maximal repetition. For k greater than the maximal repetition we have $f(k|X_1^n) = n-k+1$. If we want to have a nondecreasing function of k, which is often more convenient, we may consider $f(k|X_1^{n+k-1})$. Now let us observe that $f(k|X_1^{n+k-1}) \le \min\left[(\operatorname{card}\mathbb{X})^k, n\right]$ [19]. For a stationary process this bound can be strengthened in this way.

Theorem 1 *For a stationary process $(X_i)_{i\in\mathbb{Z}}$ we have*

$$\mathbf{E}\, f(k|X_1^{n+k-1}) \le \tilde{S}_{nk} := \sum_{w\in\mathbb{X}^k} \min\left[1, nP(w)\right] \ . \tag{4}$$

Remark Obviously, $\tilde{S}_{nk} \le \min\left[(\operatorname{card}\mathbb{X})^k, n\right]$.

Proof We have

$$f(k|X_1^{n+k-1}) = \sum_{w\in\mathbb{X}^k} \mathbf{1}\left\{\sum_{i=0}^{n-1} \mathbf{1}\left\{X_{i+1}^{i+k} = w\right\} \ge 1\right\} \ .$$

Hence by Markov inequality,

$$\begin{aligned}\mathbf{E}\, f(k|X_1^{n+k-1}) &= \sum_{w\in\mathbb{X}^k} P\left(\sum_{i=0}^{n-1} \mathbf{1}\left\{X_{i+1}^{i+k} = w\right\} \ge 1\right) \\ &\le \sum_{w\in\mathbb{X}^k} \min\left[1, \mathbf{E}\sum_{i=0}^{n-1} \mathbf{1}\left\{X_{i+1}^{i+k} = w\right\}\right] \\ &= \sum_{w\in\mathbb{X}^k} \min\left[1, nP(w)\right] \ .\end{aligned}$$

For independent random variables, bound (4) can be strengthened again. Let $o_k(f(k))$ denote a term that divided by $f(k)$ vanishes in infinity, i.e., $\lim_{k\to\infty} \frac{o_k(f(k))}{f(k)} = 0$ [15, Chap. 9].

Theorem 2 ([14, Theorem 2.1]) *For a sequence of independent identically distributed (IID) random variables $(X_i)_{i\in\mathbb{Z}}$ we have*

$$\mathbf{E}\, f(k|X_1^{n+k-1}) + o_n(1)o_k(1) = S_{nk} := \sum_{w\in\mathbb{X}^k} \left(1 - (1-P(w))^n\right) \ . \tag{5}$$

Remark We also have

$$S_{nk} = \sum_{w\in\mathbb{X}^k} P(w) \sum_{i=0}^{n-1} (1-P(w))^i \le \tilde{S}_{nk} \ .$$

Formula (5) is remarkable. It states that the expectation of subword complexity for an IID process is asymptotically such as if each substring w were drawn n times with replacement with probability $P(w)$. Correlations among overlaps of a given string asymptotically cancel out on average [14]. Function S_{nk} is also known in the analysis of large number of rare events (LNRE) [21], developed to investigate Zipf's law in quantitative linguistics [2, 30].

For the sake of further reasoning it is convenient to rewrite quantities $\tilde{S}_{nk}$ and S_{nk} as expectations of certain functions of block probability. For $x > 0$, let us define

$$\tilde{g}(x) := \min[x, 1] \ , \tag{6}$$

$$g_n(x) := x\left(1 - \left(1 - \frac{1}{nx}\right)^n\right) \ . \tag{7}$$

We also introduce

$$g(x) := \lim_{n\to\infty} g_n(x) = x\left(1 - \exp\left(-\frac{1}{x}\right)\right) \ . \tag{8}$$

Then we obtain

$$\frac{\tilde{S}_{nk}}{n} = \mathbf{E}\,\tilde{g}\left(\frac{1}{nP(X_1^k)}\right) \ , \tag{9}$$

$$\frac{S_{nk}}{n} = \mathbf{E}\,g_n\left(\frac{1}{nP(X_1^k)}\right) \approx \mathbf{E}\,g\left(\frac{1}{nP(X_1^k)}\right) \ , \tag{10}$$

where the last formula holds for sufficiently large n (looking at the graphs of g_n and g, for say $n \geq 20$).

Usually probability of a block decreases exponentially with the increasing block length. Thus it is convenient to rewrite formulae (9) and (10) further, using minus log-probability $Y_k = -\log P(X_1^k)$. The expectation of this random variable equals, by definition, block entropy $\mathbf{E}\,Y_k = H(k)$. In contrast, we obtain

$$\frac{\tilde{S}_{nk}}{n} = \mathbf{E}\,\tilde{\sigma}(Y_k - \log n) \ , \tag{11}$$

$$\frac{S_{nk}}{n} = \mathbf{E}\,\sigma_n(Y_k - \log n) \approx \mathbf{E}\,\sigma(Y_k - \log n) \ , \tag{12}$$

where $\tilde{\sigma}(y) := \tilde{g}(\exp(y))$, $\sigma_n(y) := g_n(\exp(y))$, and $\sigma(y) := g(\exp(y))$.

Apparently, formulae (4) and (5) combined with (11) and (12) could be used for estimating block entropy of a process. In fact, we have the following proposition:

Theorem 3 *For a stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$, we have*

$$\frac{\tilde{S}_{nk}}{n} = \tilde{\sigma}\left(H(k) + o_k(k) - \log n\right) + o_k(1) \ , \tag{13}$$

$$\frac{S_{nk}}{n} = \sigma_n\left(H(k) + o_k(k) - \log n\right) + o_k(1) \tag{14}$$

$$\approx \sigma\left(H(k) + o_k(k) - \log n\right) + o_k(1) \ . \tag{15}$$

Proof By the Shannon-McMillan-Breiman theorem (asymptotic equipartition property), for a stationary ergodic process, the difference $Y_k - H(k)$ is of order $o_k(k)$ with probability $1 - o_k(1)$ [1]. Since functions $\tilde{\sigma}(x)$ and $\sigma_n(x)$ increase in the considered domain and take values in range [0, 1], then the claims follow from formulae (11) and (12) respectively.

Simulations performed in the next section suggest that for a large class of processes, observed subword complexity $f(k|X_1^n)$ is practically equal to its expectation. Hence if n is large and the term $o_k(k)$ in inequality (15) is negligible, Theorems 2 and 3 suggest the following estimation procedure for block entropy of IID proceses. First, we compute the subword complexity for a sufficiently large sample X_1^n and, secondly, we apply some inverse function to obtain an estimate of entropy. Namely, by (5) and (15), we obtain the following estimate of block entropy $H(k)$,

$$H^{(1)}_{\text{est}}(k) := \log(n - k + 1) + \sigma^{-1}\left(\frac{f(k|X_1^n)}{n-k+1}\right) \ . \tag{16}$$

Formula (16) is applicable only to sufficiently small k, which stems from using sample X_1^n rather than X_1^{n+k-1}. Consequently, this substitution introduces an error that grows with k and explodes at maximal repetition. As we have mentioned, for k greater than the maximal repetition we have $f(k|X_1^n) = n - k + 1$, which implies $H^{(1)}_{\text{est}}(k) = \infty$, since $\lim_{x\to 1}\sigma^{-1}(x) = \infty$. For k smaller than the maximal repetition, we have $H^{(1)}_{\text{est}}(k) < \infty$.

Estimator $H^{(1)}_{\text{est}}(k)$ resembles in spirit inverting Padé approximant proposed by Ebeling and Pöschel [10]. Quality of this estimator will be tested empirically for Bernoulli processes in the next section. In fact, formula (16) works very well for uniform probabilities of characters. Then terms $o_k(k)$ and $o_k(1)$ in inequality (15) vanish. Thus we can estimate entropy of relatively large blocks, for which only a tiny fraction of typical blocks can be observed in the sample. Unfortunately, this estimator works so good only for uniform probabilities of characters. The term $o_k(k)$ in inequality (15) is not negligible for nonuniform probabilities of characters. The more nonuniform the probabilities are, the larger the term $o_k(k)$ is. The sign of this term also varies. It is systematically positive for small k and systematically negative for large k. Hence reliable estimation of entropy via formula (16) is impossible in general. This suggests that the approach of [10] cannot be trusted, either.

The persistent error of formula (16) comes from the fact that the asymptotic equipartition is truly only an asymptotic property. Now we will show how to improve the estimates of block entropy for quite a general stationary process. We will show that terms $o_k(k)$ and $o_k(1)$ in equality (13) may become negligible for $\tilde{S}_{nk}/n$ close to $1/2$. Introduce the Heaviside function

$$\theta(y) = \begin{cases} 0 & y < 0 \ , \\ 1/2 & y = 0 \ , \\ 1 & y > 0 \ . \end{cases} \tag{17}$$

In particular, $\mathbf{E}\,\theta(Y_k - B)$ is a decreasing function of B. Thus we can define $M(k)$, the median of minus log-probability Y_k of block X_1^k, as

$$M(k) := \sup\left\{B : \mathbf{E}\,\theta(Y_k - B) \geq 1/2\right\} \ . \tag{18}$$

Theorem 4 *For any* $C > 0$ *we have*

$$\frac{\tilde{S}_{nk}}{n} \geq \frac{1+\exp(-C)}{2} \implies M(k) \geq \log n - C \ , \tag{19}$$

$$\frac{\tilde{S}_{nk}}{n} < \frac{1}{2} \implies M(k) \leq \log n \ . \tag{20}$$

Proof First, we have $\tilde{\sigma}(y) \leq (1-\exp(-C))\theta(y+C) + \exp(-C)$. Thus

$$\tilde{S}_{nk}/n = \mathbf{E}\,\tilde{\sigma}(Y_k - \log n) \leq (1-\exp(-C))\,\mathbf{E}\,\theta(Y_k - \log n + C) + \exp(-C) \ .$$

Hence if $\tilde{S}_{nk}/n \geq (1+\exp(-C))/2$ then $\mathbf{E}\,\theta(Y_k - \log n + C) \geq 1/2$ and consequently $M(k) \geq \log n - C$. As for the converse, $\tilde{\sigma}(y) \geq \theta(y)$. Thus

$$\tilde{S}_{nk}/n = \mathbf{E}\,\tilde{\sigma}(Y_k - \log n) \geq \mathbf{E}\,\theta(Y_k - \log n + C) \ .$$

Hence if $\tilde{S}_{nk}/n < 1/2$ then $\mathbf{E}\,\theta(Y_k - \log n) < 1/2$ and consequently $M(k) \leq \log n$.

In the second step we can compare the median and the block entropy.

Theorem 5 *For a stationary ergodic process* $(X_i)_{i\in\mathbb{Z}}$*, we have*

$$M(k) = H(k) + o_k(k) \ . \tag{21}$$

Proof The claim follows by the mentioned Shannon-McMillan-Breiman theorem. Namely, the difference $Y_k - H(k)$ is of order $o_k(k)$ with probability $1 - o_k(1)$. Hence the difference $M(k) - H(k)$ must be of order $o_k(k)$ as well.

A stronger inequality may hold for a large subclass of processes. Namely, we suppose that the distribution of strings of a fixed length is skewed towards less probable values. This means that the distribution of minus log-probability Y_k is right-skewed. Consequently, those processes satisfy this condition:

Definition 1 A stationary process $(X_i)_{i\in\mathbb{Z}}$ is called properly skewed if for all $k \geq 1$ we have

$$H(k) \geq M(k) \ . \tag{22}$$

Assuming that the variance of subword complexity is small, formulae (22), (19), and (4) can be used to provide a simple lower bound of block entropy for properly skewed processes. The recipe is as follows. First, to increase precision, we extend functions $s(k)$ of natural numbers, such as subword complexity $f(k|X_1^n)$ and block entropy $H(k)$, to real arguments by linear interpolation. Namely, for $r = qk + (1-q)(k-1)$, where $q \in (0,1)$, we will put $s(r) := qs(k) + (1-q)s(k-1)$. Then we apply this procedure.

1. Choose a $C > 0$ and let $S := (1 + \exp(-C))/2$.
2. Compute $s(k) := f(k|X_1^n)/(n-k+1)$ for growing k until the least k is found such that $s(k) \geq S$. Denote this k as k_1 and let $k_2 := k_1 - 1$.
3. Define

$$k^* := \frac{(S - s(k_2))\,k_1 + (s(k_1) - S)\,k_2}{s(k_1) - s(k_2)} \ . \tag{23}$$

   (We have $s(k^*) = S$).
4. Estimate the block entropy $H(k^*)$ as

$$H^{(2)}_{\text{est}}(k^*) := \log(n - k^* + 1) - C. \tag{24}$$

5. If needed, estimate entropy rate $h = \lim_{k\to\infty} H(k)/k$ as

$$h^{(2)}_{\text{est}} = H^{(2)}_{\text{est}}(k^*)/k^*. \tag{25}$$

For a fixed sample size n, the above procedure yields an estimate of block entropy $H(k)$ but only for a single block length $k = k^*$. Thus to compute the estimates of block entropy $H(k)$ for varying k, we have to apply the above procedure for varying n. So extended procedure is quite computationally intensive and resembles baking a cake from which we eat only a cherry put on the top of it. By Theorems 1 and 4 we conjecture that estimator $H^{(2)}_{\text{est}}(k^*)$ with a large probability gives a lower bound of block entropy $H(k^*)$ for properly skewed processes. The exact quality of this bound remains, however, unknown, because we do not know the typical difference between subword complexity and function $\tilde{S}_{nk}$. Judging from the experiments with Bernoulli processes described in the next section, this difference should be small but it is only our conjecture. Moreover, it is an open question whether k^* tends to infinity for growing n and whether estimator $h_{\text{est}}^{(2)}$ is consistent, i.e., whether $h_{\text{est}}^{(2)} - h$ tends to 0 for $n \to \infty$. That the estimator $H_{\text{est}}^{(2)}(k^*)$ yields only a lower bound of block entropy is not a great problem in principle since, if some upper bound is needed, it can be computed using a universal code, such as the Lempel-Ziv code [31, 32].

3 Simulations for Bernoulli Processes

In this section we will investigate subword complexity and block entropy for a few samples drawn from Bernoulli processes. The Bernoulli(p) process is an IID process over a binary alphabet $\mathbb{X} = \{0, 1\}$ where $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. We have generated five samples X_1^n of length $n = 50000$ drawn from Bernoulli(p) processes, where $p = 0.1, 0.2, 0.3, 0.4, 0.5$. Subsequently, the empirical subword complexities $f(k|X_1^n)$ have been computed for k smaller than the maximal repetition and plotted in Fig. 1 together with the expectation S_{nk} computed from the approximate formula

$$\frac{S_{nk}}{n} = \sum_{s=0}^{k} \binom{k}{s} p^s (1-p)^{k-s} g_n \left(\frac{p^{-s}(1-p)^{-k+s}}{n} \right) \tag{26}$$

$$\approx \sum_{s=0}^{k} q_k(s) g_n \left(\frac{p^{-s}(1-p)^{-k+s}}{n} \right) , \tag{27}$$

where

$$q_k(s) = \frac{\exp\left(-\frac{(p-s/k)^2 k}{p(1-p)}\right)}{\sum_{r=0}^{k} \exp\left(-\frac{(p-r/k)^2 k}{p(1-p)}\right)} . \tag{28}$$

As we can see in Fig. 1, both the variance of $f(k|X_1^n)$ and the error term $o_n(1)o_k(1)$ in formula (5) are negligible since even for a single sample X_1^n, subword complexity $f(k|X_1^n)$ is practically equal to its expectation. We suppose that this property holds also for stochastic processes with some dependence and thus estimation of entropy from empirical subword complexity may be a promising idea.

Thus let us move on to estimation of entropy. For the Bernoulli(p) process we have block entropy $H(k) = hk$, where the entropy rate is

$$h = -p \log p - (1-p) \log(1-p) . \tag{29}$$

In Fig. 2, we compare block entropy $H(k)$ and estimator $H_{\text{est}}^{(1)}(k)$ given by formula (16). Only for p close to 0.5 estimator $H_{\text{est}}^{(1)}(k)$ provides a good estimate of block entropy $H(k)$, for block lengths smaller than roughly 25. Let us note that there are 2^{25}

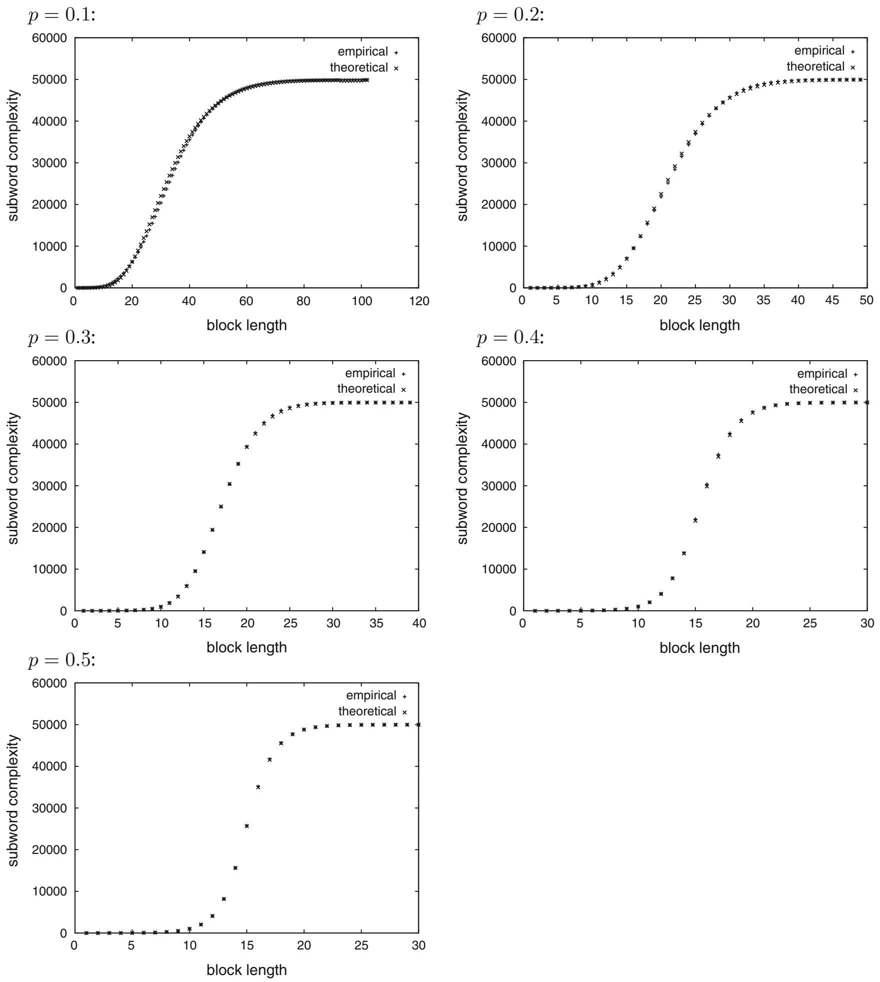


Fig. 1 Subword complexity as a function of block length for samples X_1^n drawn from Bernoulli(p) processes, where $n = 50000$ and $p = 0.1, 0.2, 0.3, 0.4, 0.5$. Pluses are the empirical data $f(k|X_1^n)$. Crosses are values S_{nk} (practically the same data points)

different blocks of length 25. This number is three orders of magnitude larger than the considered sample length ($n = 50000$). In this sample we may hence observe only a tiny fraction of the allowed blocks and yet via formula (16) we can arrive at a good estimate of block entropy. Unfortunately, the estimates $H_{\text{est}}^{(1)}(k)$ become very poor for strongly nonuniform probabilities. As we can see in Fig. 2, the sign of the difference between $H(k)$ and $H_{\text{est}}^{(1)}(k)$ varies. Moreover, whereas $H(k) = hk$ grows linearly, the shape of function $H_{\text{est}}^{(1)}(k)$ is much less regular, partly it resembles a hyperbolic function k^{β}, where $\beta \in (0, 1)$, partly it looks linear, and it is not necessarily concave.

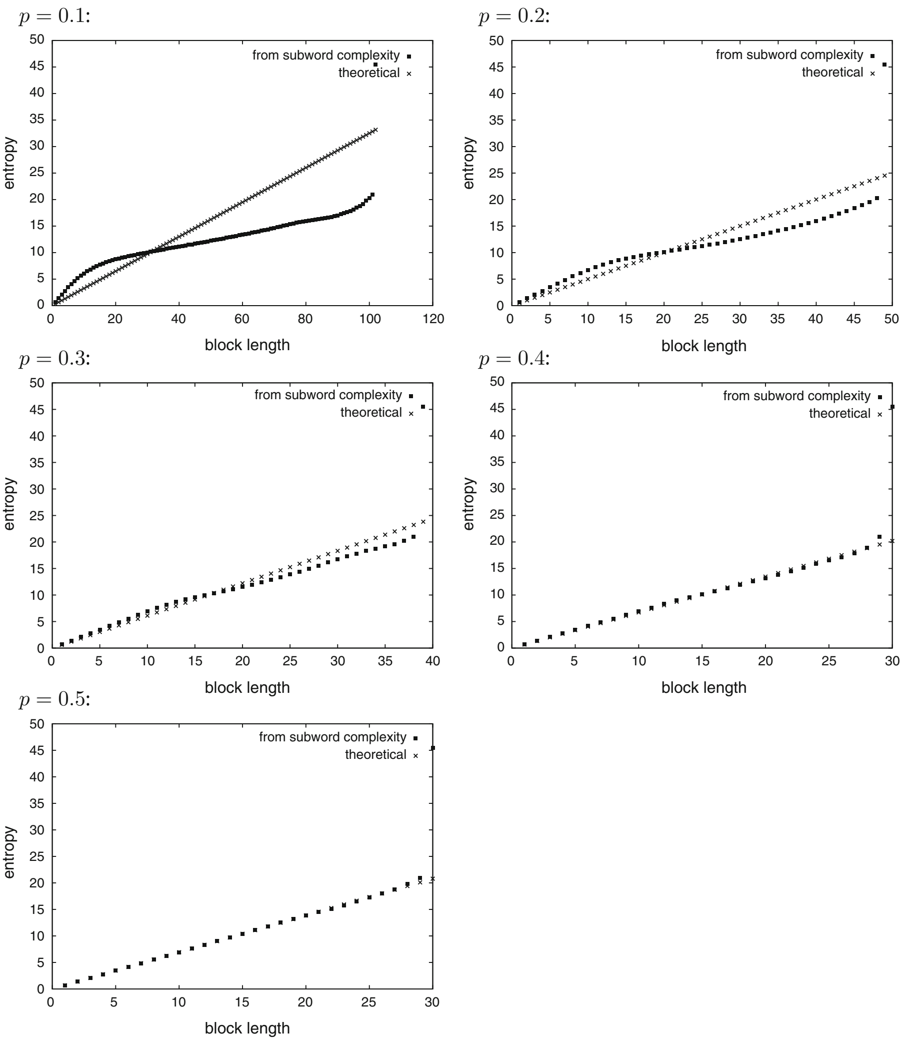


Fig. 2 Entropy as a function of block length for Bernoulli(p) processes, where $p = 0.1, 0.2, 0.3, 0.4, 0.5$. Crosses are the true values $H(k) = hk$. Squares are estimates (16) for samples X_1^n of length $n = 50000$

(Whereas true block entropy $H(k)$ is concave [4]). Hence function $H_{\text{est}}^{(1)}(k)$ cannot provide a reliable estimate of block entropy in general.

To illuminate the source of this phenomenon, in Fig. 3, empirical subword complexity $f(k|X_1^n)$ has been contrasted with function

$$f_{\text{pred}}(k|X_1^n) := (n - k + 1)\sigma\left(H(k) - \log(n - k + 1)\right) \ , \qquad (30)$$

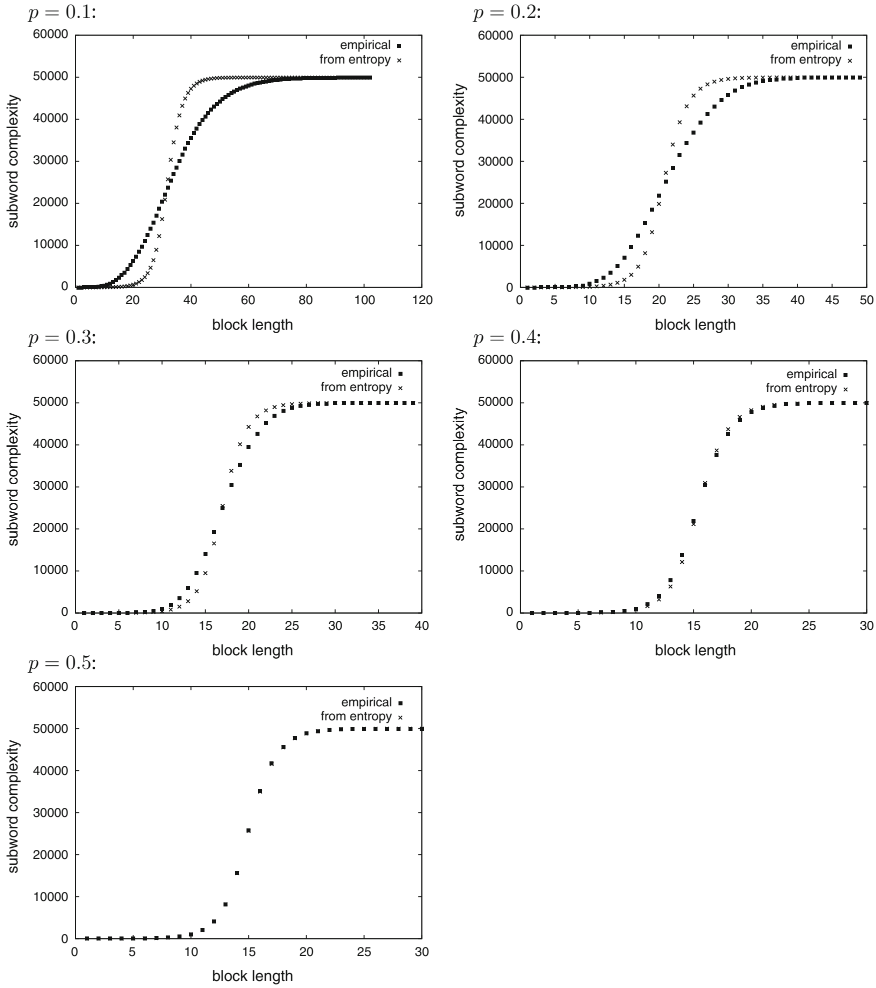


Fig. 3 Subword complexity as a function of block length for samples X_1^n drawn from Bernoulli(p) processes, where $n = 50000$ and $p = 0.1, 0.2, 0.3, 0.4, 0.5$. Squares are the empirical data $f(k|X_1^n)$. Crosses are values (30)

which should equal $f(k|X_1^n)$ if the term $o_k(k)$ in inequality (15) is negligible, n is large, and the variance of subword complexity is small. Whereas we have checked that the variance of $f(k|X_1^n)$ is small indeed, the observed difference between the empirical subword complexity and function $f_{\text{pred}}(k|X_1^n)$ must be attributed to the term $o_k(k)$ in inequality (15). As we can see in Fig. 3, the term $o_k(k)$ vanishes for uniform probabilities ($p = 0.5$) but its absolute value grows when parameter p diverges from 0.5 and can become quite substantial. The term $o_k(k)$ is systematically positive for small block lengths and systematically negative for large block lengths but it vanishes for $f(k|X_1^n) \approx n/2$.

As we have explained in the previous section, the last observation can be used to derive estimators $H_{\text{est}}^{(2)}(k)$ and $h_{\text{est}}^{(2)}$ given in formulae (24) and (25). Now we will check how these estimators work. The distribution of log-probability of blocks is approximately symmetric for Bernoulli processes, so these processes are probably properly skewed and consequently estimators $H_{\text{est}}^{(2)}(k)$ and $h_{\text{est}}^{(2)}$ should be smaller than the true values. Our simulation confirms this hypothesis. We have generated five samples X_1^m of length $m = 70000$ drawn from Bernoulli(p) processes, where $p = 0.1, 0.2, 0.3, 0.4, 0.5$. For each of these samples we have computed estimator $H_{\text{est}}^{(2)}(k)$ for $C = 2$ and subsamples X_1^n of length $n = 2^j$, where $j = 1, 2, 3, \ldots$ and $n \le m$. The results are shown in Fig. 4. As we can see in the plots, the difference between $H(k)$ and $H_{\text{est}}^{(2)}(k)$ is almost constant and close to $C = 2$. Additionally, in Fig. 5, we present the estimates of entropy rate for Bernoulli processes given by estimator $h_{\text{est}}^{(2)}$ for $C = 2$ and a sample of length $n = 50000$. They are quite rough but consistently provide a lower bound as well. For the considered sample, the relative error ranges from 17 % for uniform probabilities ($p = 0.5$) to 20 % for $p = 0.05$. Thus we may say that the performance of estimators $H_{\text{est}}^{(2)}(k)$ and $h_{\text{est}}^{(2)}$ is good, at least for Bernoulli processes.

4 Texts in Natural Language

In the previous section, we have checked that the block entropy estimator $H_{\text{est}}^{(2)}(k)$ given by formula (24) returns quite good estimates for Bernoulli processes and persistently provides a lower bound of the true block entropy. Hence in this section, we would like to apply this estimator to some real data such as texts in natural language. As we have mentioned, there were some attempts to estimate block entropy for texts in natural language from frequencies of blocks [10–12]. These attempts were quite heuristic, whereas now we have an estimator of block entropy that may work for some class of processes.

Let us recall that estimator $H_{\text{est}}^{(2)}(k)$ works under the assumption that the process is properly skewed, which holds e.g. if the distribution of strings of a fixed length is skewed towards less probable values. In fact, in texts in natural language, the empirical distribution of orthographic words, which are strings of varying length, is highly skewed in the required direction, as described by Zipf's law [30]. Hence we may suppose that the hypothetical process of generating texts in natural language is also properly skewed. Consequently, estimator $H_{\text{est}}^{(2)}(k)$ applied to natural language data should be smaller than the true block entropy.

Having this in mind, let us make some experiment with natural language data. In the following, we will analyze three texts in English: *First Folio/35 Plays* by William Shakespeare (4, 500, 500 characters), *Gulliver's Travels* by Jonathan Swift (579, 438

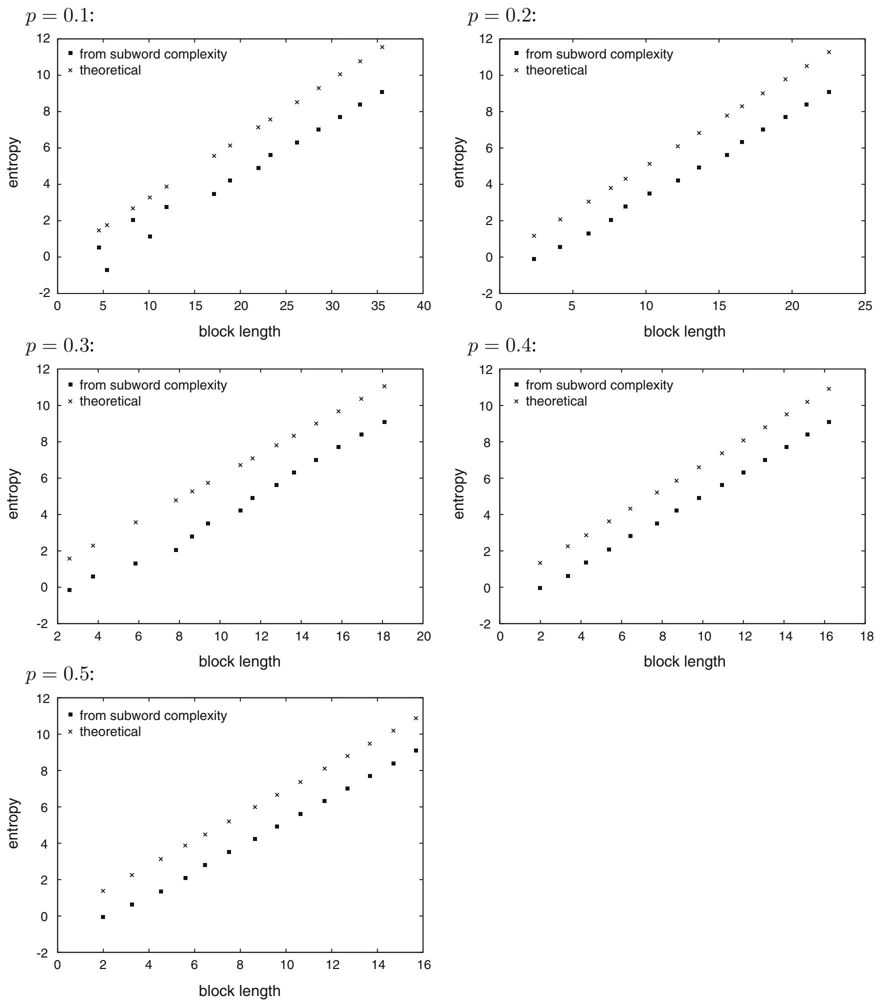


Fig. 4 Entropy as a function of block length for Bernoulli(p) processes, where $p = 0.1, 0.2, 0.3, 0.4, 0.5$. Crosses are the true values $H(k) = hk$. Squares are estimates (24) for $C = 2$ and samples X_1^n of varying length $n = 2^j$ where $n < 70000$

characters), and *Complete Memoirs* by Jacques Casanova de Seingalt (6, 719, 801 characters), all downloaded from the Project Gutenberg.[1] For each of these texts we have computed estimator $H_{\text{est}}^{(2)}(k)$ for $C = 2$ and initial fragments of texts of length $n = 2^j$, where $j = 1, 2, 3, \dots$ and n's are smaller than the text length. In this way we have obtained the data points in Fig. 6. The estimates look reasonable. The maximal block length for which the estimates can be found is $k \approx 10$, and in this

[1] www.gutenberg.org.

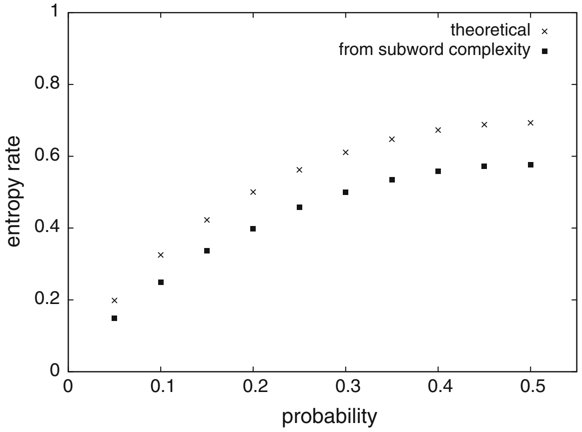


Fig. 5 Entropy rate h for Bernoulli(p) processes as a function of parameter p. Crosses are the true values, given by formula (29). Squares are the estimates given by formula (25) for $C = 2$ and samples X_1^n of length $n = 50000$

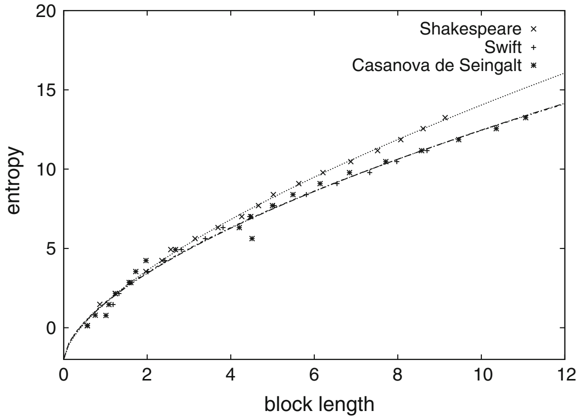


Fig. 6 Estimates of block entropy obtained through estimator (24) for $C = 2$. Crosses relate to *First Folio/35 Plays* by William Shakespeare, squares relate to *Gulliver's Travels* by Jonathan Swift, and stars relate to *Complete Memoirs* by Jacques Casanova de Seingalt. The regression functions are models (31) with $C = 2$ and the remaining parameters given in Table 1

case we obtain $H_{\text{est}}^{(2)}(k) \approx 12.5$ nats, which is less than the upper bound estimate of $H(10)/10 \approx 1.5$ nats per character by Shannon [27]. Our data also corroborate Hilberg's hypothesis. Namely, using nonlinear least squares, we have fitted model

$$H_{\text{est}}^{(2)}(k) = Ak^{\beta} - C, \tag{31}$$

where C was chosen as 2, and we have obtained quite a good fit. The values of parameters A and β with their standard errors are given in Table 1.

To verify Hilberg's hypothesis for $k \geq 10$ we need much larger data, such as balanced corpora of texts. The size of modern text corpora is of order $n = 10^9$ characters. If relationship (1) persists in so large data, then we could obtain estimates of entropy $H(k)$ for block lengths $k \leq 20$. This is still quite far from the range of data points $k \in [1, 100]$ considered by Shannon in his experiment with human subjects [27]. But we hope that estimator $H_{\text{est}}^{(2)}(k)$ can be improved to be applicable also to

Table 1 The fitted parameters of model (1). The values after ± are standard errors

Text	A	β
First Folio/35 Plays by William Shakespeare	3.57 ± 0.05	0.652 ± 0.007
Gulliver's Travels by Jonathan Swift	3.56 ± 0.07	0.608 ± 0.012
Complete Memoirs by Jacques Casanova de Seingalt	3.60 ± 0.15	0.602 ± 0.021

so large block lengths. As we have shown, for the Bernoulli model with uniform probabilities, subword complexity may convey information about block entropy for block lengths smaller than or equal the maximal repetition. For many texts in natural language, the maximal repetition is of order 100 or greater [8]. Hence we hope that, using an improved entropy estimator, we may get reasonable estimates of block entropy $H(k)$ for $k \leq 100$.

5 Conclusion

In this paper we have considered some new methods of estimating block entropy. The idea is to base inference on empirical subword complexity—a function that counts the number of distinct substrings of a given length in the sample. In an entangled form, the expectation of subword complexity carries information about the probability distribution of blocks of a given length from which information about block entropies can be extracted in some cases.

We have proposed two estimators of block entropy. The first estimator has been designed for IID processes but it has appeared that it works well only in the trivial case of uniform probabilities. Thus we have proposed a second estimator, which works for any properly skewed stationary process. This assumption is satisfied if the distribution of strings of a given length is skewed towards less probable values. It is remarkable that the second estimator with a large probability provides a lower bound of entropy, in contrast to estimators based on source coding, which give an upper bound of entropy. We stress that consistency of the second estimator remains an open problem.

Moreover, using the second estimator, we have estimated block entropy for texts in natural language and we have confirmed earlier estimates as well as Hilberg's hypothesis for block lengths $k \leq 10$. Further research is needed to provide an estimator for larger block lengths. We hope that subword complexity carries information about block entropy for block lengths smaller than or equal the maximal repetition, which would allow to estimate entropy for $k \leq 100$ in the case of natural language.

References

1. Algoet PH, Cover TM (1988) A sandwich proof of the Shannon-McMillan-Breiman theorem. Ann. Probab. 16:899–909
2. Baayen, H (2001) Word frequency distributions. Kluwer Academic Publishers, Dordrecht
3. Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York
4. Crutchfield JP, Feldman DP (2003) Regularities unseen, randomness observed: the entropy convergence hierarchy. Chaos 15:25–54
5. Dillon WR, Goldstein M (1984) Multivariate analysis: methods and appplications. Wiley, New York
6. Dębowski Ł (2011) On the vocabulary of grammar-based codes and the logical consistency of texts. IEEE Trans. Inform. Theor. 57:4589–4599
7. Dębowski Ł (2013) A preadapted universal switch distribution for testing Hilberg's conjecture (2013). http://arxiv.org/abs/1310.8511
8. Dębowski Ł (2014) Maximal repetitions in written texts: finite energy hypothesis vs. strong Hilberg conjecture (2014). http://www.ipipan.waw.pl/~ldebowsk/
9. Dębowski Ł (2014) A new universal code helps to distinguish natural language from random texts (2014). http://www.ipipan.waw.pl/~ldebowsk/
10. Ebeling W, Pöschel T (1994) Entropy and long-range correlations in literary English. Europhys. Lett. 26:241–246
11. Ebeling W, Nicolis G (1991) Entropy of symbolic sequences: the role of correlations. Europhys. Lett. 14:191–196
12. Ebeling W, Nicolis G (1992) Word frequency and entropy of symbolic sequences: a dynamical perspective. Chaos Sol. Fract. 2:635–650
13. Ferenczi S (1999) Complexity of sequences and dynamical systems. Discr. Math. 206:145–154
14. Gheorghiciuc I, Ward MD (2007) On correlation polynomials and subword complexity. Discr. Math. Theo. Comp. Sci. AH, 1–18
15. Graham RL, Knuth DE, Patashnik O (1994) Concrete mathematics, a foundation for computer science. Addison-Wiley, New York
16. Hall P, Morton SC (1993) On the estimation of entropy. Ann. Inst. Statist. Math. 45:69–88
17. Hilberg W (1990) Der bekannte Grenzwert der redundanzfreien Information in Texten—eine Fehlinterpretation der Shannonschen Experimente? Frequenz 44:243–248
18. Ivanko EE (2008) Exact approximation of average subword complexity of finite random words over finite alphabet. Trud. Inst. Mat. Meh. UrO RAN 14(4):185–189
19. Janson S, Lonardi S, Szpankowski W (2004) On average sequence complexity. Theor. Comput. Sci. 326:213–227
20. Joe H (1989) Estimation of entropy and other functionals of a multivariate density. Ann. Inst. Statist. Math. 41:683–697
21. Khmaladze E (1988) The statistical analysis of large number of rare events, Technical Report MS-R8804. Centrum voor Wiskunde en Informatica, Amsterdam
22. Kontoyiannis I, Algoet PH, Suhov YM, Wyner AJ (1998) Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. IEEE Trans. Inform. Theor. 44:1319–1327
23. Koslicki D (2011) Topological entropy of DNA sequences. Bioinformatics 27:1061–1067
24. Krzanowski W (2000) Principles of multivariate analysis. Oxford University Press, Oxford
25. de Luca A (1999) On the combinatorics of finite words. Theor. Comput. Sci. 218:13–39
26. Schmitt AO, Herzel H, Ebeling W (1993) A new method to calculate higher-order entropies from finite samples. Europhys. Lett. 23:303–309
27. Shannon C (1951) Prediction and entropy of printed English. Bell Syst. Tech. J. 30:50–64
28. Vogel H (2013) On the shape of subword complexity sequences of finite words (2013). http://arxiv.org/abs/1309.3441
29. Wyner AD, Ziv J (1989) Some asymptotic properties of entropy of a stationary ergodic data source with applications to data compression. IEEE Trans. Inform. Theor. 35:1250–1258

30. Zipf GK (1935) The Psycho-Biology of language: an introduction to Dynamic Philology. Houghton Mifflin, Boston
31. Ziv J, Lempel A (1977) A universal algorithm for sequential data compression. IEEE Trans. Inform. Theor. 23:337–343
32. Ziv J, Lempel A (1978) Compression of individual sequences via variable-rate coding. IEEE Trans. Inform. Theor. 24:530–536

Exact Rate of Convergence of Kernel-Based Classification Rule

Maik Döring, László Györfi and Harro Walk

Abstract A binary classification problem is considered, where the posteriori probability is estimated by the nonparametric kernel regression estimate with naive kernel. The excess error probability of the corresponding plug-in decision classification rule according to the error probability of the Bayes decision is studied such that the excess error probability is decomposed into approximation and estimation error. A general formula is derived for the approximation error. Under a weak margin condition and various smoothness conditions, tight upper bounds are presented on the approximation error. By a Berry-Esseen type central limit theorem a general expression for the estimation error is shown.

Keywords Lower bound · Upper bound · Classification error probability · Kernel rule · Margin condition

AMS Classification 62G10

This work was supported in part by the National Development Agency (NFÜ, Hungary) as part of the project Introduction of Cognitive Methods for UAV Collision Avoidance Using Millimeter Wave Radar, (grant no.: KMR-12-1-2012-0008).

M. Döring
Institute of Applied Mathematics and Statistics, University of Hohenheim,
Schloss Hohenheim, 70599 Stuttgart, Germany
e-mail: maik.doering@uni-hohenheim.de

L. Györfi (✉)
Department of Computer Science and Information Theory,
Budapest University of Technology and Economics,
Magyar Tudósok Körútja 2., Budapest 1117, Hungary
e-mail: gyorfi@cs.bme.hu

H. Walk
Department of Mathematics, University of Stuttgart,
Pfaffenwaldring 57, 70569 Stuttgart, Germany
e-mail: walk@mathematik.uni-stuttgart.de

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_5

1 Introduction

We consider a binary classification problem. Using a kernel estimator for the posteriori probability, the asymptotics of the error probability is examined of the corresponding plug-in classification rule. In this paper lower and upper bounds are presented on the rate of convergence of the classification error probability.

Let the feature vector X take values in $\mathbb{R}^d$ such that its distribution is denoted by μ and let the label Y be binary valued. If g is an arbitrary decision function, then its error probability is denoted by

$$L(g) = \mathbb{P}\{g(X) \neq Y\}.$$

The Bayes decision g^* minimizes the error probability. It follows that

$$g^*(x) = \begin{cases} 1 \text{ if } \eta(x) > 1/2 \\ 0 \ \text{ otherwise,} \end{cases}$$

where the posteriori probability η is given by

$$\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}.$$

Let the corresponding error probability be

$$L^* = \mathbb{P}\{g^*(X) \neq Y\}.$$

Put

$$D(x) = 2\eta(x) - 1,$$

then the Bayes decision has the following equivalent form:

$$g^*(x) = \begin{cases} 1 \text{ if } D(x) > 0 \\ 0 \ \text{ otherwise.} \end{cases}$$

In the standard model of pattern recognition, we are given training labeled samples, which are independent and identically copies of (X, Y): $(X_1, Y_1), \dots, (X_n, Y_n)$. Based on these labeled samples, one can estimate the regression function D by D_n, and the corresponding plug-in classification rule g_n derived from D_n is defined by

$$g_n(x) = \begin{cases} 1 \text{ if } D_n(x) > 0 \\ 0 \ \text{ otherwise.} \end{cases}$$

In the following our focus lies on the rate of convergence of the excess error probability $\mathbb{E}\{L(g_n)\} - L^*$. In Sect. 2 margin conditions are discussed, which measure how fast the regression function D crosses the decision boundary. A nonparametric kernel regression estimate of D is introduced and a decomposition of the excess error probability into approximation and estimation error is considered in Sect. 3. Tight upper bounds on the approximation error are shown in Sect. 4 depending on margin and smoothness conditions on D. By a Berry-Esseen type central limit theorem a general expression for the estimation error is derived in Sect. 5. Finally, some conclusions are given.

2 Margin Conditions

Given the plug-in classification rule g_n derived from D_n it follows

$$\mathbb{E}\{L(g_n)\} - L^* \le \mathbb{E}\{|D(X) - D_n(X)|\}$$

(cf. Theorem 2.2 in Devroye et al. [3]). Therefore we may get an upper bound on the rate of convergence of the excess error probability $\mathbb{E}\{L(g_n)\} - L^*$ via the L_1 rate of convergence of the corresponding regression estimation.

However, according to Sect. 6.7 in Devroye et al. [3], the classification is easier than L_1 regression function estimation, since the rate of convergence of the error probability depends on the behavior of the function D in the neighborhood of the decision boundary

$$B_0 = \{x; D(x) = 0\}. \tag{1}$$

This phenomenon has been discovered by Mammen and Tsybakov [10], Tsybakov [13], who formulated the (strong) margin condition:

- *The strong margin condition.* Assume that for all $0 < t \le 1$,

$$\mathbb{P}\{|D(X)| \le t\} \le ct^{\alpha}, \tag{2}$$

  where $\alpha > 0$ and $c > 0$.

Kohler and Krzyżak [7] introduced the weak margin condition:

- *The weak margin condition.* Assume that for all $0 < t \le 1$,

$$\mathbb{E}\left\{\mathbb{I}_{\{|D(X)| \le t\}} |D(X)|\right\} \le ct^{1+\alpha}, \tag{3}$$

  where $\mathbb{I}$ denotes the indicator function.

Obviously, the strong margin condition implies the weak margin condition:

$$\mathbb{E}\left\{\mathbb{I}_{\{|D(X)| \le t\}} |D(X)|\right\} \le \mathbb{E}\left\{\mathbb{I}_{\{|D(X)| \le t\}} t\right\} = t\mathbb{P}\{|D(X)| \le t\} \le ct \cdot t^{\alpha}.$$

The difference between the strong and weak margin condition is that, for the strong margin condition, the event

$$\{D(X) = 0\}$$

counts. One can weaken the strong margin condition (2) such that we require

$$\mathbb{P}\{0 < |D(X)| \leq t\} \leq ct^{\alpha}. \tag{4}$$

Obviously, (4) implies (3). Under some mild conditions we have that $\alpha = 1$. (Cf. Lemma 2.) The margin conditions measure how fast the probability of a t-neighborhood of the decision boundary increases with t. A large value of α corresponds to a small probability of the neighborhood of the decision boundary, which means that the probability for events far away of the decision boundary is high. Therefore, a classifier can make the right decision more easily, hence one can expect smaller errors for larger values of α.

Recently, Audibert and Tsybakov [1] proved that if the plug-in classifier g has been derived from the regression estimate $\tilde{D}$ and if D satisfies the strong margin condition, then

$$L(g) - L^* \leq \left(\int (\tilde{D}(x) - D(x))^2 \mu(dx) \right)^{\frac{1+\alpha}{2+\alpha}}. \tag{5}$$

It is easy to see that (5) holds even under weak margin condition: we have that

$$L(g) - L^* = \mathbb{E}\left\{ \mathbb{I}_{\{g(X) \neq g^*(X)\}} |D(X)| \right\} \tag{6}$$

(cf. Theorem 2.2 in Devroye et al. [3]). Let $sign(x) = 1$ for $x > 0$ and $sign(x) = -1$ for $x \leq 0$. For fixed $t_n > 0$,

$$\begin{aligned} L(g) - L^* &= \mathbb{E}\left\{ \mathbb{I}_{\{sign\,\tilde{D}(X) \neq sign\, D(X), |D(X)| \leq t_n\}} |D(X)| \right\} \\ &\quad + \mathbb{E}\left\{ \mathbb{I}_{\{sign\,\tilde{D}(X) \neq sign\, D(X), |D(X)| > t_n\}} |D(X)| \right\} \\ &\leq \mathbb{E}\left\{ \mathbb{I}_{\{|D(X)| \leq t_n\}} |D(X)| \right\} \\ &\quad + \mathbb{E}\left\{ \mathbb{I}_{\{sign\,\tilde{D}(X) \neq sign\, D(X), |\tilde{D}(X) - D(X)| > t_n\}} |\tilde{D}(X) - D(X)| \right\}, \end{aligned}$$

therefore the weak margin condition implies that

$$\begin{aligned} L(g) - L^* &\leq ct_n^{1+\alpha} + t_n \mathbb{E}\left\{ \mathbb{I}_{\{|\tilde{D}(X) - D(X)| > t_n\}} \frac{|\tilde{D}(X) - D(X)|}{t_n} \right\} \\ &\leq ct_n^{1+\alpha} + t_n \mathbb{E}\left\{ \frac{|\tilde{D}(X) - D(X)|^2}{t_n^2} \right\}. \end{aligned}$$

For the choice

$$t_n = \left(\mathbb{E}\left\{|\tilde{D}(X) - D(X)|^2\right\}\right)^{\frac{1}{2+\alpha}}$$

we get (5).

For bounding the error probability, assume, for example, that D satisfies the *Lipschitz condition*: for any $x, z \in \mathbb{R}^d$

$$|D(x) - D(z)| \leq C\|x - z\|.$$

If D is Lipschitz continuous and X is bounded then there are regression estimates such that

$$\int (D_n(x) - D(x))^2 \mu(dx) \leq c_1^2 n^{-\frac{2}{d+2}},$$

therefore (5) means that

$$L(g) - L^* \leq \left(c_1^2 n^{-\frac{2}{d+2}}\right)^{\frac{1+\alpha}{2+\alpha}} = \left(c_1^{1+\alpha} n^{-\frac{1+\alpha}{d+2}}\right)^{\frac{2}{2+\alpha}}.$$

Kohler and Krzyżak [7] proved that for the standard plug-in classification rules (partitioning, kernel, nearest neighbor) and for weak margin condition we get that the order of the upper bound can be smaller:

$$n^{-\frac{1+\alpha}{d+2}}.$$

The main aim of this paper is to show tight upper bounds on the excess error probability $\mathbb{E}\{L(g_n)\} - L^*$ of the kernel classification rule g_n.

3 Kernel Classification

We fix $x \in \mathbb{R}^d$, and, for an $h > 0$, let the (naive) kernel estimate of $D(x)$ be

$$D_{n,h}(x) = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - 1)\mathbb{I}_{\{X_i \in S_{x,h}\}}/\mu(S_{x,h}),$$

where $S_{x,h}$ denotes the sphere centered at x with radius h. Notice that $D_{n,h}$ is not a true estimate, because its denominator contains the unknown distribution μ. However, the corresponding plug-in classification rule defined below depends only on the sign of $D_{n,h}(x)$, and so μ doesn't count. The (naive) kernel classification rule is

$$g_{n,h}(x) = \begin{cases} 1 \text{ if } D_{n,h}(x) > 0 \\ 0 \text{ otherwise,} \end{cases}$$

(cf. Devroye [2], Devroye and Wagner [4], Krzyżak [8], Krzyżak and Pawlak [9]).

If D is Lipschitz continuous and X is bounded then, for the L_1 error, one has that

$$\mathbb{E}\{|D(X) - D_{n,h}(X)|\} \leq c_2 h + \frac{c_3}{\sqrt{nh^d}},$$

(cf. Györfi et al. [5]), so for the choice

$$h = n^{-\frac{1}{d+2}}, \tag{7}$$

the L_1 upper bound implies that

$$\mathbb{E}\{L(g_{n,h})\} - L^* \leq c_4 n^{-\frac{1}{d+2}}.$$

Because of (6), we have that the excess error probability of any plug-in classification rule has the following decomposition:

$$\mathbb{E}\{L(g_{n,h})\} - L^* = \mathbb{E}\left\{\int_{\{sign\, D_{n,h}(x) \neq sign\, D(x)\}} |D(x)|\mu(dx)\right\} = I_{n,h} + J_{n,h},$$

where

$$I_{n,h} = \mathbb{E}\left\{\int_{\{sign\, \bar{D}_h(x) = sign\, D_{n,h}(x) \neq sign\, D(x)\}} |D(x)|\mu(dx)\right\}$$

and

$$J_{n,h} = \mathbb{E}\left\{\int_{\{sign\, D_{n,h}(x) \neq sign\, D(x) = sign\, \bar{D}_h(x)\}} |D(x)|\mu(dx)\right\}$$

with $\bar{D}_h(x) = \mathbb{E}\{D_{n,h}(x)\}$. $I_{n,h}$ is called approximation error, while $J_{n,h}$ is the estimation error.

4 Approximation Error

First we consider the approximation error. The following proposition means that the lower bound of the approximation error is approximately the half of the upper bound. Further, it shows that the bounds of the approximation error are mainly determined by the bandwidth h.

Proposition 1

$$\left(\frac{1}{2} + o(1)\right)\bar{I}_h \leq I_{n,h} \leq \bar{I}_h,$$

where

$$\bar{I}_h = \int_{\{sign\,\bar{D}_h(x) \neq sign\,D(x)\}} |D(x)|\mu(dx).$$

Proof The upper bound is obvious. The lower bound follows from the central limit theorem, since

$$\begin{aligned}
&\mathbb{E}\left\{\int_{\{0 \geq \bar{D}_h(x), D_{n,h}(x) \leq 0 < D(x)\}} |D(x)|\mu(dx)\right\} \\
&= \int_{\{\bar{D}_h(x) \leq 0 < D(x)\}} |D(x)|\mathbb{P}\{D_{n,h}(x) \leq 0\}\mu(dx) \\
&= \int_{\{\bar{D}_h(x) \leq 0 < D(x)\}} |D(x)|\mathbb{P}\{D_{n,h}(x) - \bar{D}_h(x) \leq -\bar{D}_h(x)\}\mu(dx) \\
&\geq \int_{\{\bar{D}_h(x) \leq 0 < D(x)\}} |D(x)|\mathbb{P}\{D_{n,h}(x) - \bar{D}_h(x) \leq 0\}\mu(dx) \\
&\geq \int_{\{\bar{D}_h(x) \leq 0 < D(x)\}} |D(x)|\left(\frac{1}{2} + o(1)\right)\mu(dx),
\end{aligned}$$

where $o(1)$ is uniform in x. This can be seen, for example, using the Berry-Esseen inequality as in the proof of Proposition 2 below. The handling of remaining integral is analogous. ∎

Kohler and Krzyżak [7] bounded the rate of convergence of the excess error probability assuming that D satisfies the weak margin condition and the Lipschitz condition. Further they assume that X has a density which is bounded away from zero:

$$f(x) \geq c' > 0 \tag{8}$$

They proved that

$$\mathbb{E}\{L(g_{n,h})\} - L^* \leq c_5 h^{1+\alpha} + \frac{c_6}{nh^d} \tag{9}$$

such that, for the choice (7),

$$\mathbb{E}\{L(g_{n,h})\} - L^* \leq c_7 n^{-\frac{1+\alpha}{d+2}}.$$

In (9) the approximation error is upper bounded by $c_5 h^{1+\alpha}$. Next we show how it follows from Proposition 1. Denote by

$$B_{0,h} = \left\{x;\, \min_{z \in B_0} \|x - z\| \leq h\right\}$$

the h-neighborhood of the decision boundary B_0 defined by (1). Let λ be the Lebesgue measure and let $M^*(B_0)$ be the outer surface (Minkowski content) of the decision boundary B_0 defined by

$$M^*(B_0) = \lim_{h \downarrow 0} \frac{\lambda(B_{0,h} \setminus B_0)}{h}.$$

Lemma 1 *If D satisfies the weak margin condition and the Lipschitz condition, then*

$$\bar{I}_h \leq c_8 h^{1+\alpha}.$$

If D satisfies the Lipschitz condition, the density f of X exists, it is bounded by f_{max} and $M^*(B_0)$ is finite, then

$$\bar{I}_h \leq c_9 h^2.$$

Proof If $x \notin B_{0,h}$ then

$$sign\, \bar{D}_h(x) = sign\, D(x).$$

Therefore

$$\begin{aligned}\bar{I}_h &= \int_{\{sign\, \bar{D}_h(x) \neq sign\, D(x)\}} |D(x)|\mu(dx)\\ &= \int_{\{sign\, \bar{D}_h(x) \neq sign\, D(x), x \in B_{0,h}\}} |D(x)|\mu(dx).\end{aligned}$$

For any fixed $x \in B_{0,h}$, there is a $z_x \in B_0$ such that $\|x - z_x\| \leq h$, which together with the Lipschitz condition implies that

$$|D(x)| = |D(x) - D(z_x)| \leq Ch.$$

Thus, by the weak margin condition

$$\begin{aligned}\bar{I}_h &\leq \int_{\{|D(x)| \leq Ch, x \in B_{0,h}\}} |D(x)|\mu(dx)\\ &\leq \int_{\{|D(x)| \leq Ch\}} |D(x)|\mu(dx)\\ &\leq c(Ch)^{1+\alpha}.\end{aligned}$$

Concerning the second half of the lemma, we have that

$$\begin{aligned}\bar{I}_h &\leq \int_{\{|D(x)| \leq Ch, x \in B_{0,h}\}} |D(x)|\mu(dx)\\ &\leq Ch \int_{\{0 < |D(x)|, x \in B_{0,h}\}} 1 \mu(dx)\end{aligned}$$

$$
\begin{aligned}
&= Ch\mu\{B_{0,h} \setminus B_0\} \\
&\leq Chf_{max}\lambda\{B_{0,h} \setminus B_0\} \\
&\leq Cc_{10}h^2.
\end{aligned}
\qquad \blacksquare
$$

The technique of the second half of the previous proof implies that $\alpha = 1$.

Lemma 2 *Let D satisfies the lower Lipschitz inequality at B_0, which means a $c^* > 0$ exists, such that for all $t \in [0, 1]$ and*

$$x \notin B_{0,c^*t}$$

it follows

$$|D(x)| > t.$$

If the density f of X exists, it is bounded by f_{max}, and the outer surface $M^*(B_0)$ is finite, then the weak margin condition holds with $\alpha = 1$.

Proof We verify (4) with $\alpha = 1$.

$$
\begin{aligned}
&\mathbb{P}\{0 < |D(X)| \leq t\} \\
&= \mathbb{P}\left\{0 < |D(X)| \leq t, X \in B_{0,c^*t} \setminus B_0\right\} + \mathbb{P}\left\{0 < |D(X)| \leq t, X \notin B_{0,c^*t}\right\} \\
&\leq \mathbb{P}\left\{X \in B_{0,c^*t} \setminus B_0\right\} + \mathbb{P}\{0 < |D(X)| \leq t, t < |D(X)|\} \\
&\leq c_{10}c^*t.
\end{aligned}
\qquad \blacksquare
$$

Hall and Kang [6] investigated the bandwidth choice. They assumed that conditional densities exist, which are bounded away from zero. Under twice differentiable conditional densities, they proved that

$$\mathbb{E}\{L(g_{n,h})\} - L^* \leq c_{11}h^4 + o\left(\frac{1}{nh^d}\right). \tag{10}$$

In (10) the approximation error is upper bounded by $c_{11}h^4$. Next we show how it follows from Proposition 1.

Let us introduce some notations:

$$p_+ := \mathbb{P}\{Y = 1\},\ p_- := \mathbb{P}\{Y = 0\}$$

Assume that the density f of X exists. Let the conditional densities f_+ and f_- be defined by

$$\mathbb{P}\{X \in A \mid Y = 1\} = \int_A f_+(x)\, dx$$

and

$$\mathbb{P}\{X \in A \mid Y = 0\} = \int_A f_-(x)\, dx.$$

Then

$$f(x) = p_+ \cdot f_+(x) + p_- \cdot f_-(x)$$

and

$$D(x) = \frac{\tilde{f}(x)}{f(x)},$$

where

$$\tilde{f}(x) := p_+ \cdot f_+(x) - p_- \cdot f_-(x).$$

Moreover,

$$f_+(x) = \frac{f(x) \cdot \big(1 + D(x)\big)}{2p_+}, \; f_-(x) = \frac{f(x) \cdot \big(1 - D(x)\big)}{2p_-}.$$

Lemma 3 *Assume that $\tilde{f}$ is two-times differentiable with bounded second order partial derivatives. If D satisfies the weak margin condition and the density f is bounded below by f_{min}, then*

$$\bar{I}_h \leq c_{12} h^{2(1+\alpha)}.$$

Proof Let $H_{\tilde{f}}$ be the Hessian-matrix of $\tilde{f}$. Then the conditions of the lemma imply that

$$\sup_{0 \leq t \leq 1} \left|(x-z)^T H_{\tilde{f}}\big(x + t(z-x)\big)(x-z)\right| \leq c_{13} \|x - z\|^2 \tag{11}$$

with $0 < c_{13} < \infty$. We have the decomposition

$$\begin{aligned}
\bar{D}_h(x) &= \frac{\int_{S_{x,h}} D(z)\mu(dz)}{\mu(S_{x,h})} \\
&= \frac{\int_{S_{x,h}} \frac{\tilde{f}(z)}{f(z)} f(z)dz}{\mu(S_{x,h})} \\
&= \frac{\int_{S_{x,h}} \Big(\tilde{f}(z) - \tilde{f}(x)\Big)\, dz}{\mu(S_{x,h})} + \frac{\tilde{f}(x)\lambda(S_{x,h})}{\mu(S_{x,h})} \\
&= \frac{\int_{S_{x,h}} \Big(\tilde{f}(z) - \tilde{f}(x)\Big)\, dz}{\mu(S_{x,h})} + D(x)\frac{f(x)\lambda(S_{x,h})}{\mu(S_{x,h})}.
\end{aligned}$$

The second order Taylor expansion

$$\tilde{f}(z) - \tilde{f}(x) = (z-x)^T \nabla \tilde{f}(x) + (z-x)^T H_{\tilde{f}}(\tilde{x}_z)(z-x)/2$$

with $\tilde{x}_z \in S_{x,h}$ implies that

$$\begin{aligned}
&\bar{D}_h(x) \\
&= \frac{\int_{S_{x,h}} \left((z-x)^T \nabla\tilde{f}(x) + (z-x)^T H_{\tilde{f}}(\tilde{x}_z)(z-x)/2\right)\, dz}{\mu(S_{x,h})} + D(x)\frac{f(x)\lambda(S_{x,h})}{\mu(S_{x,h})} \\
&= \frac{\int_{S_{x,h}} (z-x)^T H_{\tilde{f}}(\tilde{x}_z)(z-x)/2\, dz}{\mu(S_{x,h})} + D(x)\frac{f(x)\lambda(S_{x,h})}{\mu(S_{x,h})}.
\end{aligned}$$

Therefore, from (11) we get that

$$\bar{D}_h(x) \geq -\frac{c_{13}h^2\lambda(S_{x,h})/2}{\mu(S_{x,h})} + D(x)\frac{f(x)\lambda(S_{x,h})}{\mu(S_{x,h})}.$$

Thus, for $D(x) \geq 0 > \bar{D}_h(x)$, we have

$$|D(x)| \leq \frac{c_{13}h^2}{2f(x)} \leq \frac{c_{13}h^2}{2f_{min}}.$$

The same inequality holds for $D(x) < 0 \leq \bar{D}_h(x)$. From the weak margin condition we get

$$\begin{aligned}
\bar{I}_h &= \int_{\left\{sign\left(\bar{D}_h(x)\right)\neq sign\left(D(x)\right), x\in B_{0,h}\right\}} |D(x)|\mu(dx) \\
&\leq \int_{\left\{|D(x)|\leq \frac{c_{13}h^2}{2f_{min}}\right\}} |D(x)|\mu(dx) \\
&\leq c_{12}h^{2(1+\alpha)}.
\end{aligned}$$

■

Under the assumption of Lemma 2 we have that the weak margin condition holds with $\alpha = 1$. Hence by Lemma 3 and Proposition 1 we get that the approximation error $I_{n,h}$ is upper bounded by a multiple of h^4.

The question left is that whether the upper bounds in Lemmas 1 and 3 are tight. Consider some examples, where $\alpha = 1$ and $\bar{I}_h$ can be calculated showing that the order of the lower bounds have the order of the upper bounds.

Example 1 Assume that f is the uniform density on $[-1, 1]^d$. Let $h < 1$, $\beta \geq 1$, $a > 0$, $b > 0$, $a + b < 1$. Choose

$$p_+ = \frac{1}{2} + \frac{b}{4(\beta+1)}$$

and

$$D(x) = ax_1 + bx_1^\beta \cdot \mathbb{I}_{(0,1]}(x_1),$$

where $x = (x_1, \dots, x_d)$. Then

$$\begin{aligned}
\tilde{f}(x) &= \frac{ax_1 + bx_1^{\beta} \cdot \mathbb{I}_{(0,1]}(x_1)}{2^d} \cdot \mathbb{I}_{[-1,1]^d}(x) \\
f_+(x) &= \frac{1 + ax_1 + bx_1^{\beta} \cdot \mathbb{I}_{(0,1]}(x_1)}{2^{d+1}p_+} \cdot \mathbb{I}_{[-1,1]^d}(x) \\
f_-(x) &= \frac{1 - ax_1 - bx_1^{\beta} \cdot \mathbb{I}_{(0,1]}(x_1)}{2^{d+1}p_-} \cdot \mathbb{I}_{[-1,1]^d}(x).
\end{aligned}$$

One can check that D satisfies the weak margin condition with $\alpha = 1$. If $x_1 > 0$ then $sign\,\bar{D}_h(x) = sign\,D(x)$. Let V_d be the volume of the d-dimensional unit sphere, i.e. $V_d = \pi^{d/2}/\,\Gamma(d/2+1)$. For $-h < x_1 \le 0$

$$\begin{aligned}
&\int_{S_{x,h}} D(x)\;\mu(dx) = \int_{S_{x,h}} \frac{ax_1 + a(z_1 - x_1) + bz_1^{\beta} \cdot \mathbb{I}_{(0,1]}(z_1)}{2^d} \cdot \mathbb{I}_{[-1,1]^d}(z)\;dz \\
&= \frac{aV_d}{2^d}h^d x_1 + \int_{x_1-h}^{x_1+h} \frac{a(z_1 - x_1) + bz_1^{\beta} \cdot \mathbb{I}_{(0,1]}(z_1)}{2^d} V_{d-1}\big(h^2 - (z_1 - x_1)^2\big)^{(d-1)/2} dz_1 \\
&= \frac{aV_d}{2^d}h^d x_1 + \frac{bV_{d-1}}{2^d}\int_0^{x_1+h} z_1^{\beta}\big(h^2 - (z_1 - x_1)^2\big)^{(d-1)/2}\;dz_1 \\
&\le \frac{aV_d}{2^d}h^d x_1 + \frac{bV_{d-1}}{2^d}h^{d-1}\int_0^{x_1+h} z_1^{\beta}\;dz_1 \\
&\le \frac{h^{d-1}}{2^d}\Big(aV_d h x_1 + \frac{bV_{d-1}}{\beta+1}h^{\beta+1}\Big).
\end{aligned}$$

And we have a lower bound by

$$\begin{aligned}
&\int_{S_{x,h}} D(x)\mu(dx) = \frac{aV_d}{2^d}h^d x_1 + \frac{bV_{d-1}}{2^d}\int_0^{x_1+h} z_1^{\beta}\big(h^2 - (z_1 - x_1)^2\big)^{(d-1)/2}\;dz_1 \\
&= \frac{aV_d}{2^d}h^d x_1 + \frac{bV_{d-1}}{2^d}\int_0^{h} \mathbb{I}_{(-x_1,\infty)}(\tilde{z}_1)(\tilde{z}_1 + x_1)^{\beta}\big(h^2 - \tilde{z}_1^2\big)^{(d-1)/2}\;d\tilde{z}_1 \\
&= \frac{aV_d}{2^d}h^d x_1 + \frac{bV_{d-1}}{2^d}\int_0^{h} \Big(\mathbb{I}_{(0,\infty)}(\tilde{z}_1)\tilde{z}_1^{\beta}\big(h^2 - \tilde{z}_1^2\big)^{(d-1)/2} \\
&\qquad\qquad + x_1\mathbb{I}_{(-\tilde{x}_1,\infty)}(\tilde{z}_1)\beta(\tilde{z}_1 + \tilde{x}_1)^{\beta-1}\big(h^2 - \tilde{z}_1^2\big)^{(d-1)/2}\Big)\;d\tilde{z}_1 \\
&\ge \frac{aV_d}{2^d}h^d x_1 + \frac{bV_{d-1}}{2^d}\int_0^{h} \Big(\tilde{z}_1^{\beta}\big(h^2 - \tilde{z}_1^2\big)^{(d-1)/2} + x_1\beta\tilde{z}_1^{\beta-1}h^{d-1}\Big)\;d\tilde{z}_1
\end{aligned}$$

$$\geq \frac{aV_d}{2^d} h^d x_1 + \frac{bV_{d-1}}{2^d} \Big(\int_0^{h/2} \tilde{z}_1^{\beta} \big(h^2 - (h/2)^2\big)^{(d-1)/2} \, d\tilde{z}_1 + \int_{h/2}^{h} (h/2)^{\beta-1} \tilde{z}_1 \big(h^2 - \tilde{z}_1^2\big)^{(d-1)/2} \, d\tilde{z}_1 + x_1 h^d \Big)$$

$$= \frac{aV_d}{2^d} h^d x_1 + \frac{bV_{d-1}}{2^d} \Big(\frac{1}{\beta+1} (h/2)^{\beta+1} \big(h^2 - (h/2)^2\big)^{(d-1)/2} + (h/2)^{\beta-1} \big(h^2 - (h/2)^2\big)^{(d+1)/2} \frac{1}{d+1} + x_1 h^d \Big)$$

$$= \frac{h^{d-1}}{2^d} \Big((aV_d + bV_{d-1}) h x_1 + bV_{d-1} \big(\frac{(3/4)^{(d-1)/2}}{(\beta+1)2^{\beta+1}} + \frac{(3/4)^{(d+1)/2}}{(d+1)2^{\beta-1}} \big) h^{\beta+1} \Big).$$

For the notations

$$c_{14} = \frac{bV_{d-1}}{aV_d + bV_{d-1}} \big(\frac{(3/4)^{(d-1)/2}}{(\beta+1)2^{\beta+1}} + \frac{(3/4)^{(d+1)/2}}{(d+1)2^{\beta-1}} \big)$$

$$c_{15} = \frac{bV_{d-1}}{aV_d \cdot (\beta+1)}$$

we get that

$$-c_{14} h^{\beta} < x_1 \quad \Longrightarrow \quad \bar{D}_h(x) > 0 \quad \Longrightarrow \quad -c_{15} \cdot h^{\beta} < x_1$$

Therefore

$$\bar{I}_h = \int_{\{sign\, \bar{D}_h(x) \neq sign\, D(x)\}} |D(x)| \, \mu(dx)$$

$$\geq \int_{-c_{14} \cdot h^{\beta}}^{0} \int_{[-1,1]^{d-1}} -\frac{a}{2^d} x_1 \, d(x_2, \ldots x_d) \, dx_1 = \frac{ac_{14}^2}{4} \cdot h^{2\beta}$$

Analogously

$$\frac{ac_{14}^2}{4} \cdot h^{2\beta} \leq \bar{I}_h \leq \frac{ac_{15}^2}{4} \cdot h^{2\beta}$$

- If $\beta = 1$, then $D, \tilde{f}, f_+$ and f_- are Lipschitz continuous and

$$\bar{I}_h \geq c_{16} h^2.$$

- If $\beta = 1 + \epsilon/2$ for $\epsilon > 0$, then $D, \tilde{f}, f_+$ and f_- are continuously differentiable and

$$\bar{I}_h \geq c_{17} h^{2+\epsilon}.$$

- If $\beta = 2 + \epsilon/2$ for $\epsilon > 0$, then $D, \tilde{f}, f_+$ and f_- are two times continuously differentiable and
$$\bar{I}_h \geq c_{18} h^{4+\epsilon}.$$

5 Estimation Error

Next we consider the estimation error. Introduce the notations

$$N_{x,h} = \frac{\mu(S_{x,h})\bar{D}_h(x)^2}{1 - \mu(S_{x,h})\bar{D}_h(x)^2}$$

and
$$R_{x,h} = \frac{c_{19}}{\sqrt{\mu(S_{x,h})(1 - \mu(S_{x,h})\bar{D}_h(x)^2)^3}}$$

with a universal constant $c_{19} > 0$. Put

$$\bar{J}_{n,h} = \int_{\{sign\,\bar{D}_h(x) = sign\,D(x)\}} |D(x)| \Phi\left(-\sqrt{n \cdot N_{x,h}}\right) \mu(dx),$$

where Φ stands for the standard Gaussian distribution function.

Proposition 2 *We have that*

$$|J_{n,h} - \bar{J}_{n,h}| \leq \int_{\{sign\,\bar{D}_h(x) = sign\,D(x)\}} \frac{R_{x,h} \cdot |D(x)|}{\sqrt{n} + n^2 N_{x,h}^{3/2}} \mu(dx).$$

Put $\varepsilon > 0$. If the density of X exists then, for h small enough,

$$\begin{aligned}
&\int_{\{sign\,\bar{D}_h(x) = sign\,D(x)\}} |D(x)| \Phi\left(-\sqrt{(1+\varepsilon)n \cdot \mu(S_{x,h})} |\bar{D}_h(x)|\right) \mu(dx) \\
&\leq \bar{J}_{n,h} \\
&= \int_{\{sign\,\bar{D}_h(x) = sign\,D(x)\}} |D(x)| \Phi\left(-\sqrt{n \cdot \mu(S_{x,h})} |\bar{D}_h(x)|\right) \mu(dx),
\end{aligned}$$

and

$$\begin{aligned}
&\int_{\{sign\,\bar{D}_h(x) = sign\,D(x)\}} \frac{R_{x,h} \cdot |D(x)|}{\sqrt{n} + n^2 N_{x,h}^{3/2}} \mu(dx) \\
&\leq \int_{\{sign\,\bar{D}_h(x) = sign\,D(x)\}} \frac{2c_{19} \cdot |D(x)|}{\sqrt{n\mu(S_{x,h})}(1 + (\sqrt{n \cdot \mu(S_{x,h})} |\bar{D}_h(x)|)^3)} \mu(dx)
\end{aligned}$$

with a universal constant $c_{19} > 0$.

Proof First we show the following: For fixed x and h, under $0 < \bar{D}_h(x)$ we have that

$$|\mathbb{P}\{D_{n,h}(x) \leq 0\} - \Phi\left(-\sqrt{n \cdot N_{x,h}}\right)| \leq \frac{R_{x,h}}{\sqrt{n} + n^2 N_{x,h}^{3/2}},$$

which implies the first half of the proposition. (The case $\bar{D}_h(x) \leq 0$ and $D_{n,h}(x) > 0$ is completely analogous.) Introduce the notation

$$Z_i = -(2Y_i - 1)\mathbb{I}_{\{X_i \in S_{x,h}\}}.$$

Then

$$\mathbb{P}\{D_{n,h}(x) \leq 0\} = \mathbb{P}\left\{\sum_{i=1}^{n} Z_i \geq 0\right\} = \mathbb{P}\left\{\frac{\sum_{i=1}^{n}(Z_i - \mathbb{E}\{Z_i\})}{\sqrt{n\mathbb{V}ar(Z_1)}} \geq -\frac{\sqrt{n}\mathbb{E}\{Z_1\}}{\sqrt{\mathbb{V}ar(Z_1)}}\right\}.$$

Because of

$$\mathbb{V}ar(Z_1) = \mathbb{E}\{|Z_1|^2\} - (\mathbb{E}\{Z_1\})^2 = \mu(S_{x,h}) - \mu(S_{x,h})^2 \bar{D}_h(x)^2$$

and by $0 < \bar{D}_h(x)$ we have that

$$\frac{\mathbb{E}\{Z_1\}}{\sqrt{\mathbb{V}ar(Z_1)}} = -\frac{\sqrt{\mu(S_{x,h})}\bar{D}_h(x)}{\sqrt{1 - \mu(S_{x,h})\bar{D}_h(x)^2}} = -\sqrt{N_{x,h}}$$

Therefore the central limit theorem for the probability $\mathbb{P}\{D_{n,h}(x) \leq 0\}$ implies that

$$\mathbb{P}\{D_{n,h}(x) \leq 0\} = \mathbb{P}\left\{-\frac{\sum_{i=1}^{n}(Z_i - \mathbb{E}\{Z_i\})}{\sqrt{n\mathbb{V}ar(Z_1)}} \leq -\sqrt{nN_{x,h}}\right\} \approx \Phi\left(-\sqrt{nN_{x,h}}\right).$$

Notice that it is only an approximation. In order to make bounds out of the normal approximation, we refer to Berry-Esseen type central limit theorem (see Theorem 14 in Petrov [12]). Thus,

$$\left|\mathbb{P}\{D_{n,h}(x) \leq 0\} - \Phi\left(-\sqrt{nN_{x,h}}\right)\right| \leq \frac{c_{19}\frac{\mathbb{E}\{|Z_1|^3\}}{\mathbb{V}ar(Z_1)^{3/2}}}{\sqrt{n}\left(1 + \left(\sqrt{nN_{x,h}}\right)^3\right)},$$

with the universal constant $30.84 \geq c_{19} > 0$ (cf. Michel [11]). Because of $\mathbb{E}\{|Z_1|^3\} = \mu(S_{x,h})$ we get that

$$c_{19}\frac{\mathbb{E}\{|Z_1|^3\}}{\mathbb{V}ar(Z_1)^{3/2}} = \frac{c_{19}}{\mu(S_{x,h})^{1/2}\left(1-\mu(S_{x,h})\bar{D}_h(x)^2\right)^{3/2}} = R_{x,h},$$

hence

$$\left|\mathbb{P}\{D_{n,h}(x) \leq 0\} - \Phi\left(-\sqrt{nN_{x,h}}\right)\right| \leq \frac{R_{x,h}}{\sqrt{n}(1+(n \cdot N_{x,h})^{3/2})}.$$

Concerning the second half of the proposition notice that if the density of X exists then

$$R_{x,h} \leq \frac{2c_{19}}{\sqrt{\mu(S_{x,h})}},$$

and, for any $\varepsilon > 0$,

$$\mu(S_{x,h})\bar{D}_h(x)^2 \leq N_{x,h} \leq (1+\varepsilon)\mu(S_{x,h})\bar{D}_h(x)^2$$

if h is small enough. ∎

Next we show that the upper bound of the error term in the second half in Proposition 2 is of order $o\left(\frac{1}{nh_n^d}\right)$.

Lemma 4 *Assume that*

$$\lim_{n\to\infty} h_n = 0 \quad \textit{and} \quad \lim_{n\to\infty} nh_n^d = \infty, \tag{12}$$

and that there is a $\tilde{c} > 0$ *such that* $\mathit{sign}\,\bar{D}_h(x) = \mathit{sign}\,D(x)$ *implies* $|\bar{D}_h(x)| \geq \tilde{c}|D(x)|$. *If* X *has a density* f *with bounded support then*

$$A_n := \int_{\{sign\,\bar{D}_{h_n}(x)=sign\,D(x)\}} \frac{|D(x)|}{\sqrt{n\mu(S_{x,h_n})}(1+(\sqrt{n\mu(S_{x,h_n})}|\bar{D}_{h_n}(x)|)^3)}\mu(dx)$$
$$= o\left(\frac{1}{nh_n^d}\right). \tag{13}$$

Proof Let B be the bounded support of f. Then $f(x) > 0$ on B and $f(x) = 0$ on B^c. Introduce the notation

$$f_n(x) = \frac{\mu(S_{x,h_n})}{\lambda(S_{x,h_n})},$$

where λ stands for the Lebesgue measure. Under the conditions of the lemma we have that

$$\begin{aligned}
A_n &\le \int \frac{|D(x)|}{\sqrt{n\mu(S_{x,h_n})}(1+(\sqrt{n\mu(S_{x,h_n})}\tilde{c}|D(x)|)^3)}\mu(dx) \\
&= \int \frac{|D(x)|}{\sqrt{n\lambda(S_{x,h_n})f_n(x)}(1+(\sqrt{n\lambda(S_{x,h_n})f_n(x)}\tilde{c}|D(x)|)^3)}\mu(dx) \\
&= \int \frac{1}{f_n(x)}\frac{1}{\tilde{c}n\lambda(S_{x,h_n})}\frac{\sqrt{n\lambda(S_{x,h_n})f_n(x)}\tilde{c}|D(x)|}{1+(\sqrt{n\lambda(S_{x,h_n})f_n(x)}\tilde{c}|D(x)|)^3}\mu(dx) \\
&= \frac{1}{\tilde{c}nh_n^d V_d}\int_B \frac{f(x)}{f_n(x)}r_n(x)dx
\end{aligned}$$

with

$$r_n(x) = \frac{\sqrt{n\lambda(S_{x,h_n})f_n(x)}\tilde{c}|D(x)|}{1+(\sqrt{n\lambda(S_{x,h_n})f_n(x)}\tilde{c}|D(x)|)^3}.$$

Thus, we have to show that

$$\int_B \frac{f(x)}{f_n(x)}r_n(x)dx \to 0.$$

By the Lebesgue density theorem $f_n(x) \to f(x)$ and therefore $f_n(x)/f(x) \to 1$ λ—a.e. on B, and so $r_n(x) \to 0$ λ—a.e. on B. Moreover, this convergence is dominated:

$$r_n(x) \le \max_{0\le u} \frac{u}{1+u^3} =: r_{max}.$$

Thus,

$$\int_B r_n(x)dx \to 0.$$

Apply the decomposition

$$\begin{aligned}
\int_B \frac{f(x)}{f_n(x)}r_n(x)dx &\le \int_B \left|\frac{f(x)}{f_n(x)}-1\right| r_n(x)dx + \int_B r_n(x)dx \\
&\le r_{max}\int_B \left|\frac{f(x)}{f_n(x)}-1\right| dx + o(1).
\end{aligned}$$

In order to prove

$$\int_B \left|\frac{f(x)}{f_n(x)}-1\right| dx \to 0 \tag{14}$$

we refer to the Riesz-Vitali-Scheffé theorem, according to which (14) is satisfied if

$$\frac{f}{f_n} \ge 0,$$

$$\frac{f}{f_n} \to 1 \quad \lambda\text{—a.e. on } B,$$

and

$$\int_B \frac{f(x)}{f_n(x)} dx \to \int_B 1\, dx = \lambda(B). \tag{15}$$

Thus, it remains to show (15). By the generalized Lebesgue density theorem (cf. Lemma 24.5 in Györfi et al. [5]), for each μ-integrable function m

$$\int_B \left| \frac{\int_{S_{x,h_n}} m(z)\mu(dz)}{\mu(S_{x,h_n})} - m(x) \right| \mu(dx) \to 0.$$

Therefore

$$\int_B \frac{\int_{S_{x,h_n}} m(z)\mu(dz)}{\mu(S_{x,h_n})} \mu(dx) \to \int_B m(x)\mu(dx).$$

Choose

$$m(x) = \frac{1}{f(x)},\ x \in B.$$

Then

$$\int_B \frac{f(x)}{f_n(x)} dx = \int_B \frac{\int_{S_{x,h_n}} m(z)\mu(dz)}{\mu(S_{x,h_n})} \mu(dx) \to \int_B m(x)\mu(dx) = \int_B 1\, dx = \lambda(B),$$

and the lemma is proved. ∎

As we already mentioned, using Hoeffding and Bernstein inequalities Kohler and Krzyżak [7] proved that under the condition (8) we have

$$J_{n,h_n} \le \frac{c_6}{nh_n^d} \tag{16}$$

with $c_6 < \infty$.

We believe that applying Proposition 2 the condition (8) can be weakened such that the following conjecture holds: If X is bounded and it has a density then we have (16).

Because of Lemma 4, this conjecture means that

$$\int_{\{sign\, \bar{D}_h(x) = sign\, D(x)\}} |D(x)| \Phi\left(-\sqrt{n\mu(S_{x,h_n})}|\bar{D}_{h_n}(x)|\right) \mu(dx) \le \frac{c_6}{nh_n^d}. \tag{17}$$

Concerning a possible way to prove (17) we may apply the covering argument of (5.1) in Györfi et al. [5], which says that for bounded X,

$$\int \frac{1}{\mu(S_{x,h_n})}\mu(dx) \leq \frac{c_{20}}{h_n^d}.$$

The bounded support of X can be covered by spheres $S_{x_j,h_n/2}, j = 1, \dots, M_n$ such that $M_n \leq c_{21}/h_n^d$. Let

$$S'_{x,h_n} = S_{x,h_n} \cap \{z : \ sign\, \bar{D}_{h_n}(z) = sign\, D(z)\}.$$

If $x \in S'_{x_j,h_n/2}$ then $S'_{x_j,h_n/2} \subseteq S'_{x,h_n}$. For (17),

$$\begin{aligned}
&\int_{\{sign\, \bar{D}_{h_n}(x) = sign\, D(x)\}} |D(x)| \Phi\left(-\sqrt{n\mu(S_{x,h_n})}|\bar{D}_{h_n}(x)|\right)\mu(dx)\\
&\leq \sum_{j=1}^{M_n} \int_{S'_{x_j,h_n/2}} |D(x)| \Phi\left(-\sqrt{n\mu(S_{x,h_n})}|\bar{D}_{h_n}(x)|\right)\mu(dx)\\
&\leq \sum_{j=1}^{M_n} \int_{S'_{x_j,h_n/2}} |D(x)| \Phi\left(-\sqrt{n\mu(S_{x_j,h_n/2})}|\bar{D}_{h_n}(x)|\right)\mu(dx)\\
&\leq \frac{1}{n}\sum_{j=1}^{M_n} \int_{S'_{x_j,h_n/2}} n|D(x)| e^{-n\mu(S_{x_j,h_n/2})|\bar{D}_{h_n}(x)|^2/2}\mu(dx)\\
&\leq \frac{1}{n}\sum_{j=1}^{M_n} \int_{S'_{x_j,h_n/2}} n|D(x)| e^{-n\mu(S_{x_j,h_n/2})\tilde{c}^2|D(x)|^2/2}\mu(dx),
\end{aligned}$$

where the last inequality follows by the assumptions of Lemma 4. If

$$\sup_j \int_{S'_{x_j,h_n/2}} n|D(x)| e^{-n\mu(S_{x_j,h_n/2})\tilde{c}^2|D(x)|^2/2}\mu(dx) < \infty$$

then

$$\int_{\{sign\, \bar{D}_h(x) = sign\, D(x)\}} |D(x)| \Phi\left(-\sqrt{n\mu(S_{x,h_n})}|\bar{D}_{h_n}(x)|\right)\mu(dx) \leq c_{22}\frac{M_n}{n} \leq \frac{c_6}{nh_n^d}.$$

Example 2 Notice that the upper bound in (16) is tight. As in the Example 1, if μ is the uniform distribution and

$$D(x) = x_1,$$

then $sign\, \bar{D}_{h_n}(x) = sign\, D(x)$ and $|\bar{D}_{h_n}(x)| = |D(x)|$ for $h_n < 1/2$ and $|x_1| < 1/2$. Thus

$$
\begin{aligned}
&J_{n,h_n}\\
&\geq \int_{\{sign\,\bar{D}_{h_n}(x)=sign\,D(x)\}} |D(x)|\Phi\left(-\sqrt{(1+\varepsilon)n\mu(S_{x,h_n})}|\bar{D}_{h_n}(x)|\right)\mu(dx)\\
&\geq \int_0^{1/2} z\Phi\left(-\sqrt{(1+\varepsilon)V_d n h_n^d 2^{-d}}z\right)dz\\
&= \frac{1}{(1+\varepsilon)V_d n h_n^d 2^{-d}}\int_0^{\sqrt{(1+\varepsilon)V_d n h_n^d 2^{-d}}/2} u\Phi(-u)\,du\\
&= \frac{1}{(1+\varepsilon)V_d n h_n^d 2^{-d}}\left(\int_0^{\infty} u\Phi(-u)\,du + o(1)\right)\\
&\geq \frac{c_{23}}{nh_n^d},
\end{aligned}
$$

with $0 < c_{23}$. Hall and Kang [6] proved that if X has a density, bounded from above and from below then

$$J_{n,h_n} = o\left(\frac{1}{nh_n^d}\right),$$

which contradicts the lower bound $\frac{c_{23}}{nh_n^d}$.

In general, we conjecture the following: If X has a density, which is bounded by f_{max}, then

$$\frac{c_{24}}{nh_n^d} \leq J_{n,h_n}$$

with $0 < c_{24}$. This conjecture is supported by the fact that

$$\mu(S_{x,h_n}) \leq f_{max} V_d h_n^d.$$

Therefore

$$
\begin{aligned}
&\int_{\{sign\,\bar{D}_{h_n}(x)=sign\,D(x)\}} |D(x)|\Phi\left(-\sqrt{(1+\varepsilon)n\mu(S_{x,h_n})}|\bar{D}_{h_n}(x)|\right)\mu(dx)\\
&\geq \int_{\{sign\,\bar{D}_{h_n}(x)=sign\,D(x)\}} |D(x)|\Phi\left(-\sqrt{(1+\varepsilon)nf_{max}V_d h_n^d}|\bar{D}_{h_n}(x)|\right)\mu(dx).
\end{aligned}
$$

6 Conclusion

We presented tight upper bounds for the rate of convergence of the error probability of kernel classification rule. Decomposing the excess error probability into the sum of approximation and estimation error, we derived approximate formulas both for approximation error and estimation error.

Under weak margin condition with α and Lipschitz condition on the regression function D, Kohler and Krzyżak [7] showed that the approximation error $I_{n,h}$ is upper bounded by $c_5 h^{\alpha+1}$. If, in addition, the conditional densities are twice continuously differentiable, then we proved that $I_{n,h} \leq c_{12} h^{2(\alpha+1)}$. Furthermore, we present an example, according to which these upper bounds are tight. Under the assumption that the Minkowski content of the decision boundary is finite and the lower Lipschitz inequality holds, the weak margin condition holds with $\alpha = 1$. Hence we get the upper bound $I_{n,h} \leq c_{11} h^4$ as in Hall and Kang [6] as a special case.

If X has a lower bounded density, then Kohler and Krzyżak proved the upper bound on the estimation error: $J_{n,h} \leq c_6/(nh^d)$. We show that this upper bound is tight, too.

References

1. Audibert J-Y, Tsybakov AB (2007) Fast learning rates for plug-in classifiers, Ann Stat 35:608–633
2. Devroye L (1981) On the almost everywhere convergence of nonparametric regression function estimates. Ann Stat 9:1310–1319
3. Devroye L, Györfi L, Lugosi G (1996) A probabilistic theory of pattern recognition. Springer, New York
4. Devroye L, Wagner TJ (1980) Distribution-free consistency results in nonparametric discrimination and regression function estimation. Ann Stat 8:231–239
5. Györfi L, Kohler M, Krzyżak A, Walk H (2002) A distribution-free theory of nonparametric regression. Springer, New York
6. Hall P, Kang K (2005) Bandwidth choice for nonparametric classification. Ann Stat 33:284–306
7. Kohler M, Krzyżak A (2007) On the rate of convergence of local averaging plug-in classification rules under a margin condition. IEEE Trans Inf Theory 53:1735–1742
8. Krzyżak A (1986) The rates of convergence of kernel regression estimates and classification rules. IEEE Trans Inf Theory IT-32:668–679
9. Krzyżak A, Pawlak M (1984) Distribution-free consistency of a nonparametric kernel regression estimate and classification. IEEE Trans Inf Theory IT-30:78–81
10. Mammen E, Tsybakov AB (1999) Smooth discrimination analysis. Ann Stat 27:1808–1829
11. Michel R (1981) On the constant in the non-uniform version of the Berry-Esseen theorem. Z Wahrsch Verw Gebiete 55:109–117
12. Petrov VV (1975) Sums of independent random variables. Springer, Berlin
13. Tsybakov AB (2004) Optimal aggregation of classifiers in statistical learning. Ann Stat 32:135–166

Compound Bipolar Queries: A Step Towards an Enhanced Human Consistency and Human Friendliness

Janusz Kacprzyk and Sławomir Zadrożny

Abstract Database querying is a basic capability to make use of databases that are omnipresent and huge. A crucial problem is how to make possible for an ordinary human user to properly express his intentions and preferences as to what should be searched for. As natural language, with its inherent imprecision, is the only fully natural human means of communication and articulation, this makes difficult the use of traditional binary logic based querying tools. Fuzzy logic can come to the rescue, notably using fuzzy logic with linguistic quantifiers. Such queries, proposed by Kacprzyk and Ziółkowski [24], Kacprzyk et al. [25], have offered much in this context, and will also be used here. While looking for further solutions in this direction, the concept of a bipolar query has been proposed by Dubois and Prade [13], followed by a fuzzy bipolar query due to Zadrożny [36] (cf. Zadrożny and Kacprzyk [40]) involving negative and positive information, notably meant as required and desired conditions. A natural solution consisting in combining these two ideas was proposed conceptually by Kacprzyk and Zadrożny [22], and termed a *compound bipolar query*. In this paper we further extend this concept mainly by exploring some additional aggregation related aspects of bipolar queries which include fuzzy queries with linguistic quantifiers.

To Professor Jacek Koronacki, a great researcher and scholar, and a long time friend, for his wise support and inspiration.

J. Kacprzyk (✉) · S. Zadrożny
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

S. Zadrożny
e-mail: Slawomir.Zadrozny@ibspan.waw.pl

J. Kacprzyk
WIT—Warsaw School of Information Technology, Warsaw, Poland

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_6

1 Introduction

The purpose of this short note is to discuss some possible approaches, within the line of research we have been active in for many years, to the very essence of *flexible querying*, that is retrieving from (possibly large) numerical databases information which the human agent really wants to find. The human agent is assumed to be an average user, not a database specialist, and for him or her the only natural way of articulation of his or her intentions or needs, but also preferences (i.e. which record is good and which is not), is by using natural language a characteristic feature of which is an inherent imprecision and vagueness.

Needless to say that the use of natural language per se is a great difficulty to the traditional database querying systems and languages but this is not all. Namely, the "normal", or average human users think—as the human beings do—about a more sophisticated and complex concepts corresponding to what they want to retrieve. For instance, they may prefer to speak about a "good financial standing" rather than about, for instance, "yearly income is much more than USD xx" and "loans in banks are negligible" etc.

However, those concepts, being consistent for humans, are difficult to represent formally through some commands of a querying language, notably the SQL. The problem is clearly not an inherent imprecision of terms like a "high income" because they can be represented by using fuzzy logic, and this has been known since the late 1960s (cf. a survey in Galindo [18]). The very problem is a proper representation of the very meaning of those concepts which are at a higher level of abstraction like a "good financial standing". Clearly, a proper representation should involve some aggregation of satisfaction values from the fulfillment of some conditions on some attributes but it is not clear which type of aggregation should be employed. The usual AND-type aggregation (and its related OR-type) has been dealt with in the traditional fuzzy querying literature by just using the minimum (or a t-norm) or maximum (or an s-norm) but this was not enough.

The first step towards a more realistic and human consistent approach to the representation of more complex concepts in flexible (fuzzy) querying was proposed in the mid-1986 by Kacprzyk and Ziółkowski [24], followed by Kacprzyk et al. [25] in which the concept of a *database query with a fuzzy linguistic quantifier* has been introduced.

The idea of query with a linguistic quantifier is as follows. For clarity we will quote here an example from the real world application for which these queries have been initially proposed at the International Institute for Applied Systems Analysis in Laxenburg, Austria. The queries have been supposed to be employed in database querying to be used in a decision support system for regional authorities responsible for, among other issues, water quality. Those top level people, unfamiliar with databases, have—while being interviewed during the development of a conceptual solution and architecture of that decision support system—expressed interest mainly in highly aggregated and sophisticated concepts like "dangerous water pollution". After some discussion it has turned out that their human perception of the very meaning of such concepts has been that, for instance, "most of the important water pollution indicators have considerably exceeded limits", with the linguistic quantifier "most" considered to be very important.

It is easy to see that meaning of that complex term, "dangerous water pollution", equated with the above linguistically quantified statement, has been a challenge, mostly because of the occurrence of the linguistic quantifier "most" that has been outside of the scope of the traditional logic, and its very essence was even not the same as the essence of the *generalized quantifiers* (cf. Mostowski [29] or Lindström [28]). Luckily enough, Zadeh's [35] fuzzy logic based calculus of linguistically quantified propositions makes it possible to determine the degrees of truth of such linguistically quantified statements. A very important capability has therefore been made available by fuzzy querying with linguistic quantifiers.

However, as powerful as fuzzy queries with linguistic quantifiers can be, they have not been in a position to express all fine shades of meaning of real human intentions as to what they really want to retrieve. Among many possible steps in this context, one in which we have been active for years (cf. Zadrożny [36], Zadrożny and Kacprzyk [40], Zadrożny et al. [37], just to name a few) has been the concept of a *bipolar query*.

The very essence of a *bipolar query* can be best shown on the following simple example of looking for a "good" apartment in a database of a real estate agency. Suppose that a customer comes to an agent and, when asked about what he or she wants to find, responds that he or she is interested in the purchasing of a "good apartment". As such a general intention cannot be directly related to database attributes and used by the agent to search the database, the customer is again asked what he or she means by a good apartment. The response may often be that he or she looks for an "inexpensive apartment" *and possibly*, "close to public transportation". Notice, however, that this is still far from the traditional pattern of queries because "and possibly" is certainly not equivalent to the classical conjunction.

The above type of a database query may be viewed as a reflection of the so called *bipolarity*. Namely, there is a sound evidence, resulting from many experiments in social and cognitive sciences, that a human being is usually considering both positive and negative information while making a judgment or decision. This positive and negative information is then somehow aggregated yielding an overall judgment. The way this aggregation is carried out may be more or less explicit but in general it may take different forms, depending on the task, context, emotional involvement etc. In the literature, often a special interpretation of the positive and negative information is assumed which assigns higher importance to the latter but in a rather subtle way. Basically, considered in the database querying context, the latter implies mandatory requirements, as negative information is treated as corresponding to what must not occur, and the former implies optional requirements, as positive information is treated as concerning what is preferable, if possible. This *bipolarity* should therefore be reflected in database queries.

In our context, formal and algorithmic attempts at the formalization of bipolarity in judgments and evidence are crucial, and for us in particular those based on fuzzy logic and possibility theory the roots of which constitute fundamental works by

Dubois and Prade and their collaborators, to name just a few: Benferhat et al. [4] or Dubois and Prade [14–16].

There are many aspects of bipolarity to be taken into account, cf. Zadrożny et al. [37]. First, various models of bipolarity may be adopted, notably the use of a proper scale to describe the phenomenon. Basically, two such models (scales) are usually considered: *bipolar univariate* and *unipolar bivariate*, cf. Grabisch et al. [19]. The former assumes one scale with three main levels of, respectively, negative, neutral and positive evaluation, gradually changing, while the latter model assumes two more or less independent scales which separately account for a positive and negative evaluation. Usually, the intervals $[-1, 1]$ and $[0, 1]$ are used to represent the scales in the respective models of bipolarity. We will use the unipolar bivariate scales to handle bipolarity.

It is easy to see that bipolarity triggers a qualitatively different character of bipolar evaluations. In flexible querying, a crucial problem is how to use them to order the resulting set of tuples. Basically, one should aggregate them and obtain an overall evaluation of each tuple against a bipolar query but, in general, it is not clear how to do this.

Here we study a special case when positive and negative conditions may be interpreted in terms of desired and required conditions. Then, their aggregation may proceed as proposed for the crisp case by Lacroix and Lavency [26]. We follow this approach, adapt it to the fuzzy case and study its properties and possible implementations.

Basically, a prototypical example of a bipolar query considered in this paper is

$$C \text{ and possibly } P \tag{1}$$

exemplified by

find a house which is *inexpensive* **and possibly** *close to public transportation*

The main problem is to find a proper aggregation method to reflect the very tricky and non-conventional aggregation operator "…and possibly,…". In our analyses we will adopt our approach to that aggregation (cf. Zadrożny [36], Zadrożny and Kacprzyk [38, 40]), De Tré et al. [11, 12]) that is basically an extension of the classic Lacroix and Lavency [26] (nonfuzzy) approach to queries with preferences.

The bipolar queries as in the example given under (1) do well serve their purpose as a means for a much more human consistent way of expressing what the human being may want to retrieve (which house to find) but clearly represent a simple situation, like a first approximation, of how real intentions and preferences are formulated by the human being.

In addition to the choice of a model of bipolarity, as mentioned earlier, one has also to decide how to combine bipolar evaluations of elementary conditions comprising the whole query and how to order query results with respect to bipolar evaluations.

The first approach may be traced to the seminal work of Lacroix and Lavency [26] on queries with preferences composed of two types of conditions: those which are required to be met and those which are just desired to be met. Such a query may be exemplified as follows: "Find houses cheaper than USD 500 K and located not more than five blocks from a railway station". Data not satisfying the former condition (i.e., houses more expensive than USD 500 K) are readily rejected, while the dissatisfaction of the latter condition (i.e., located farther than five blocks from the station) may be acceptable provided there do not exist houses satisfying both conditions simultaneously. Thus the former conditions may be treated as providing negative information, related to the mandatory requirements pointing out which features of data makes them totally uninteresting to the user. The latter conditions may be seen as providing positive information, related to the optional requirements pointing out which features of data make them really interesting to the user. In that approach we adopt the unipolar bivariate model of bipolarity, and we basically assume that the bipolarity appears on the level of the whole query. Our main concern is clearly how to combine (aggregate) both the negative and positive evaluations to obtain a total evaluation on a traditional unipolar univariate scale that would provide simple means for ordering the results of the query.

The introduction of a bipolar query has been a huge step forward towards a higher human consistency and human friendliness of flexible (fuzzy) query but it has not solved all problems.

Basically, though a bipolar query reflects the way the human beings perceive how their real intention as to what is sought is formalized, the two parts of a bipolar query, in its simple form concern real attributes in the database, e.g., the "price" and "distance to a railway station". Such an approach is justified because many formal properties can be formulated and proved (cf. the recent Zadrożny and Kacprzyk's [40] paper). This involvement of real database attributes may be, however, too strong a limitation in many real situations—cf. Kacprzyk and Zadrożny [22, 23].

Namely, in most cases, a customer looking for a house or apartment uses in the beginning much more general form of his or her intention, by stating, if we stay in the real estate context, for instance

> find a house that is *affordable* **and possibly** *well located*

which can be considered a zero (highest) level query formulation (intention/preference)—cf. Kacprzyk and Zadrożny [22].

Then, if a real (human) estate agent is involved, then he or she would certainly ask to elaborate on the mandatory/necessary (later on referred to as: *required*) and optional/possible (later on referred to as: *desired*) conditions. And then, for instance, the customer states

- for the required condition ("an affordable house"), I mean

  > find a house that is *inexpensive* **and possibly** in a *modern building*

- for the desired condition ("well located"), I mean

  > find a house that is *in an affluent part of the town* **and possibly** *close to a recreational area*

and these can be viewed as the first level query formulations (intentions/preferences).

One can clearly continue in the same manner and obtain, by asking the customer to further elaborate on his or her intentions, for instance

- for the first level required condition, in the required part, i.e. "inexpensive"

  find a house that has a *low price* **and possibly** a *good bank loan offer*

- for the first level required condition, in the desired part, i.e. "modern building"

  find a house that is in a building with an *intelligent energy management* **and possibly** with *fast lifts*

- for the first level optional formulation, in the required part, i.e. "an affluent part of the town"

  find a house that is in a *quiet zone* **and possibly** is *close to the business district*

- for the first level desired formulation, in the desired part, i.e. "close to a recreational area"

  find a house that is close to a *park* **and possibly** *close to a lake*

and, obviously, one can continue in the similar way by further extending the particular parts of the queries.

This is an example of what can happen if the customer is inclined to stay within the bipolar query context, that is, if his intentions at consecutive levels can be best expressed by bipolar queries.

This need not always be the case.

Our point of departure is here again that those imprecisely specified desired and required conditions are clearly not directly related to the particular attributes in the database, so that they cannot directly be employed for querying. Usually, those conditions may adequately be represented by some aggregation of conditions on attribute values. For instance, such an aggregation can proceed by using a linguistic quantifier driven aggregation via Zadeh's [35] calculus of linguistically quantified propositions. To be more specific, combination of the desired and required condition may itself be a condition of a fuzzy query. For convenience, we will present this new extension in terms of our fuzzy querying interface, FQUERY for Access, which supports an extended version of SQL, cf. Kacprzyk and Zadrożny [20, 21] and Bosc and Pivert [6] in the sense that, first, it supports *linguistic terms* in queries exemplified by fuzzy values such as "inexpensive" and fuzzy relations (fuzzy comparison operators) such as "much less than" as in the following SQL query

```
SELECT *
FROM   apartment
WHERE  (price IS inexpensive) AND                (2)
       (distance to bus stop IS
       muchlessthan 4 blocks)
```

So, to best express the customer's intention, we can use in the required and desired conditions of the source bipolar query *linguistic quantifiers* such as "most", "almost all" etc. which play the role of flexible aggregation operators. This leads to the concept of a *bipolar fuzzy query with a linguistic quantifier*.

Therefore, staying for clarity in the real estate context, we have the following situation

- we assume a bipolar query of the general type

$$C \text{ and possibly } P \tag{3}$$

  in which we have

  - a required condition C, exemplified in the real estate case considered by "affordable",
  - a desired condition P, exemplified in the real estate case by "well located",

- we assume that the required and desired conditions involve a linguistic quantifier driven aggregation of some partial conditions (which directly correspond to attributes in a real estate database in question!).

This boils down to the determination of a degree of truth of a linguistically quantified proposition in the sense of Zadeh [35].

For instance, suppose that

- the required condition "affordable" is defined as follows

$$Q \text{ of conditions among } \{c_i\}_{i=1,\dots,n_C} \text{ are to be satisfied} \tag{4}$$

  where n_C is the number of conditions in C, exemplified by

  > *Most* of conditions among 'price IS *inexpensive*, bank loan IS *easy to get*, bank interest IS *not much higher than x%*, …' are to be satisfied

- the desired condition "conveniently located" is defined as follows

$$Q \text{ of conditions among } \{p_i\}_{i=1,\dots,n_P} \text{ are to be satisfied} \tag{5}$$

  where n_P is the number of conditions in P, exemplified by

  > *Most* of conditions among 'distance to railroad station IS *short*, distance to bus stop IS *much less than 5 blocks*, number of buses at stop IS *high*, …' are to be satisfied

Therefore, in the new so called compound bipolar queries proposed Kacprzyk and Zadrożny [23] we have a traditional bipolar query, denoted as (C, P), in which both C and P are fuzzy queries themselves, namely, fuzzy queries with linguistic quantifiers.

We will now present first a more detailed exposition of the compound fuzzy query proposed. We will start with a brief summary of the essence of fuzzy queries with linguistic quantifiers and then of bipolar queries in Sects. 2 and 3, respectively. Finally, we will present the essence of our new proposal combining the bipolar queries with queries with linguistic quantifiers.

2 Queries with Linguistic Quantifiers

We assume that the linguistic quantifiers are meant in the sense of Zadeh, and can be handled by two basic calculi: Zadeh's original calculus based on fuzzy logic and by using Yager's OWA operators. We will show the use of the former.

Zadeh's calculus of linguistically quantified propositions concerns natural language type expressions like

$$\text{“}\textit{Most}\text{ Swedes are }\textit{tall}\text{”} \tag{6}$$

where “Most” is an example of a linguistic quantifier. Other examples include “almost all”, “much more than 50 %” etc. These are examples of so-called *relative* quantifiers we are interested in.

A linguistically quantified proposition exemplified by (6) may be written as

$$Qx\ A(x) \tag{7}$$

where Q denotes a linguistic quantifier (e.g., *most*), $X = \{x\}$ is a universe of discourse (e.g., a set of Swedes), and $A(x)$ is a predicate corresponding to a certain property (e.g., of being *tall*).

The problem is to compute the truth value of (7). First, Q is equated with a fuzzy set defined in [0, 1], and we assume that Q is a regular nondecreasing quantifier, such that

$$y_1 \leq y_2 \Rightarrow \mu_Q(y_1) \leq \mu_Q(y_2); \quad \mu_Q(0) = 0; \quad \mu_Q(1) = 1 \tag{8}$$

where the particular $y \in [0, 1]$ correspond to proportions of elements with property A and $\mu_Q(y)$ is the degree to which a given proportion matches the semantics of Q. For example, $Q =$ “most” might be given as

$$\mu_Q(y) = \begin{cases} 1 & \text{for } y > 0.8 \\ 2y - 0.6 & \text{for } 0.3 \leq y \leq 0.8 \\ 0 & \text{for } y < 0.3 \end{cases} \tag{9}$$

The predicate A is modeled by an appropriate fuzzy set A defined in X with its membership function μ_A.

Formally, the truth degree of (7) is computed using the following formula

$$\text{Truth}(Qx\ A(x)) = \mu_Q(\frac{1}{n}\sum_{i=1}^{n}\mu_A(x_i)) \tag{10}$$

where n is the cardinality of X.

In the case of linguistically quantified statements with importance, exemplified by "*Most* (Q) of *young* (B) Swedes are *tall* (A)", written generally as

$$QBx\ A(x) \tag{11}$$

where Q denotes a linguistic quantifier (e.g., *most*), $X = \{x\}$ is a universe of discourse (e.g., a set of Swedes), $B(x)$ denotes importance of the particular x's, and $A(x)$ is a predicate corresponding to a certain property (e.g., of being *tall*), we have

$$\text{Truth}(QBx\ A(x)) = \mu_Q(\frac{\sum_{i=1}^{n}\mu_B(x_i)\wedge\mu_A(x_i)}{\sum_{i=1}^{n}\mu_B(x_i)})$$

where "$\wedge$" is the minimum but may be replaced by a t-norm, if appropriate.

To find Truth$(Qx\ A(x))$ and Truth$(QBx\ A(x))$ one can also use Yager's *ordered weighted averaging (OWA) operators* defined as follows. Let $W \in [0,1]^m$, $W = [w_1,\ldots,w_m]$, $\sum_{i=1}^{m} w_i = 1$ be a weight vector. Then the OWA operator of dimension m and weight vector W is a function $O_W : [0,1]^m \longrightarrow [0,1]$ such that

$$O_W(a_1,\ldots,a_m) = W \circ B = \sum_{i=1}^{m} w_i b_i \tag{12}$$

where b_i is ith largest element among a_i's and $B = [b_1,\ldots,b_m]$; $\circ$ denotes the scalar product.

The OWA operators generalize many widely used aggregation operators. In particular one obtains the maximum, minimum and average operators assuming $W = [1,0,\ldots,0,0]$, $W = [0,0,\ldots,0,1]$ and $W = [\frac{1}{m},\frac{1}{m},\ldots,\frac{1}{m},\frac{1}{m}]$, respectively.

The OWA operators may be used to model linguistic quantifiers. Assume that Q is a regular non-decreasing linguistic quantifier in the sense of Zadeh (8). Then the weight vector of the corresponding OWA operator is defined by Yager [30] as follows

$$w_i = \mu_Q\left(\frac{i}{m}\right) - \mu_Q\left(\frac{i-1}{m}\right), \qquad i = 1,\ldots,m \tag{13}$$

Then, we can easily find the corresponding truth values of linguistically quantified statements.

Since the purpose of this paper is to introduce a new class of compound bipolar queries involving as a required and desired condition a query with linguistic quantifiers, we will only assume that the truth value of a linguistically quantified proposition representing the query is calculated by using some method, and the type of the method (mainly Zadeh's [35] calculus of linguistically quantified propositions or Yager's [30] OWA operators) does not matter for our further considerations.

The fuzzy linguistic quantifiers may occur in various clauses of the SQL query and we will follow the approach adopted in our FQUERY for Access package, cf. Kacprzyk and Zadrożny [20, 21], i.e., they will be used as operators aggregating conditions in the `WHERE` clause. For example, with "most" defined by (9), the interpretation of that aggregation may be

$$\text{“}\textit{Most}\text{ of the predicates } \{A_i\}_{i=1,\ldots,n} \text{ are satisfied”} \tag{14}$$

for any number, n, of predicates.

Therefore, we have formal means to define and implement queries with fuzzy linguistic quantifiers, denoted for simplicity as

- $Qx\ A(x)$ for the query with a linguistic quantifier of type (7), the truth value of which is given by (4), i.e.

$$\text{Truth}(QxP(x)) = \mu_Q(\frac{1}{n}\sum_{i=1}^{n}\mu_P(x_i))$$

- $QBx\ A(x)$ for the query with a linguistic quantifier of type (11), the truth value of which is given by (12), i.e.

$$\begin{aligned}&\text{Truth}(QBxP(x))\\&= \mu_Q(\frac{\sum_{i=1}^{n}\mu_B(x_i)\wedge\mu_P(x_i)}{\sum_{i=1}^{n}\mu_B(x_i)})\end{aligned}$$

where "$\wedge$" is the minimum but may be replaced by another t-norm.

3 Bipolar Fuzzy Database Queries

Now we will briefly show the concept of a bipolar fuzzy query as meant in this paper and discuss some properties and related aspects that will be of relevance for this paper, referring the reader to our other publications for details, notably Zadrożny and Kacprzyk [40].

As we have already mentioned, a query is to be meant as a condition expressing what the user is looking for, and the response to a query is a list of tuples satisfying this condition(s). For simplicity we assume just for now that we have simple, atomic conditions with constraints on the values of the attributes characterizing a given relation (table), and these atomic conditions are connected using the logical connectives of the conjunction, disjunction and negation; notice that we do not consider for now the use of linguistic quantifiers as proposed by Kacprzyk and Ziółkowski [24], Kacprzyk et al. [25]; and they will be introduced into the bipolar queries later in this paper.

Basically, such a simple fuzzy query concerning attribute X using a linguistic term modeled by fuzzy set A may be denoted as

$$X \text{ is } A \tag{15}$$

where X in (15) may denote the attribute `price`, while the fuzzy set A may represent the linguistic term "low".

A *unipolar scale* is clearly associated with (15) as $\mu_A(x)$ denotes the degree to which a given attribute value is compatible with the meaning of a given linguistic term and hence the degree to which this value satisfies the query condition. Therefore, this is a *unipolar fuzzy querying* approach.

Bipolarity in (fuzzy) database querying is essentially meant as the incorporation of *negative* and *positive* assessments/evaluations of data. In our real estate context, for a customer the location near a railroad station may be welcome (a positive assessment) while its high price is discouraging (a negative assessment). Even more, the placement near a station may be attractive (positively assessed) due to a commuting convenience and at the same time unattractive (negatively assessed) due to, e.g., the noise and social environment.

The first aspect, as already mentioned in the Introduction, that is crucial is related to a proper scale for expressing the bipolarity. An effective and efficient solution is to employ two types of scales (cf. Grabisch et al. [19]) *bipolar univariate* and *unipolar bivariate*. In the former, an assessment is expressed as a number from $[-1, 1]$, which is divided into three parts expressing the negative ($<$0), neutral (0) and positive ($>$0) assessments, respectively. In the latter, the positive and negative assessments are expressed separately on two unipolar scales with values from [0, 1]. In our case it will be more convenient to use two separate unipolar scales what is effective and efficient; cf. for instance Yorke [34] for some psychological justifications.

Another important aspect is the level of data to which assessments are applied. They can refer to: the values of the particular attribute domains, and to the whole tuples (cf. [12]).

Bipolar queries have been first conceptualized in the context of fuzzy logic and using its tools and techniques by Dubois and Prade [13] (cf. Dubois and Prade [14] for a comprehensive exposition) and their basic idea is to distinguish two types of query conditions which are related to the *(flexible) constraints* representing what is required (this, or better to say its negation, corresponds to the negative condition) and what is merely *desired* (this directly corresponds to the positive condition). Thus, in Dubois and Prade's approach there is no symmetry between negative and positive evaluations: the former are treated as more important. This is clear, in fact by their very semantics. This is confirmed by the original Dubois and Prade's strategy to generate an answer to such a query which may be briefly stated as "first select (with respect to the negative condition) and then order (with respect to the positive condition)". To implement this strategy for fuzzy conditions, Dubois and Prade [13, 14, 16] propose to employ the lexicographic order of the tuples represented by vectors of two degrees of matching of the negative and positive conditions. This is clearly a legitimate, yet conceptually simple and effective and efficient solution.

Research on bipolar queries in the sense adopted in this paper was triggered by the seminal paper of Lacroix and Lavency [26] who proposed the use of a query with two categories of conditions (similarly to the approach of Dubois and Prade, mentioned above): C which is required (mandatory), and P which expresses just mere preferences (desires). Such a query obviously involves bivariate unipolar scale with the negative evaluation corresponding to *not* satisfying the required condition C, and positive evaluation corresponding directly to satisfying the desired condition P. Thus, concerning the very interpretation of negative and positive evaluations it is the same approach which was later adopted by Dubois and Prade in the above mentioned works, as well as by current authors in virtually all their works on bipolar queries.

The crucial aspect of the semantics of Lacroix and Lavency's approach is related to the "and possibly" operator in (1). Such an aggregation operator has been later proposed independently by Dubois and Prade in default reasoning and earlier by Yager [31, 32] in multicriteria decision making with so-called *possibilistically qualified criteria* which should be satisfied unless they interfere with the satisfaction of other criteria. This is in fact the essence of the aggregation operator "and possibly" as proposed by Lacroix and Lavency. This concept was also applied by Bordogna and Pasi [5] in information retrieval.

Lacroix and Lavency [26] consider only the case of crisp conditions C and P so that a bipolar query (C, P) may be processed using the "first select using C then order using P" strategy, i.e., the answer to the bipolar query (C, P) is obtained by, first, finding tuples satisfying C and, second, choosing from among them those satisfying P, if any. Such a very meaning was also assumed while fuzzifying the Lacroix and Lavency approach, notably in Zadrożny [36], and Zadrożny and Kacprzyk [38], and is used also here. As to other approaches, cf. Bosc and Pivert [7, 8], Dubois and Prade [14] or Lietard et al. [27].

Suppose that queries are addressed to a set of tuples $T = \{t_j\}$ (a relation). The negative and positive assessments defining a bipolar query are identified with the predicates (fuzzy sets) that represent them, denoted as C and P, respectively (the negative assessment corresponds to the complement of C). For a tuple $t \in T$, $C(t)$ and $P(t)$ denote either that t satisfies the respective condition (in crisp case) or a degree of this satisfaction, in a fuzzy case. Therefore, the whole bipolar query is denoted, as mentioned earlier, by (C, P).

Lacroix and Lavency's [26] aggregation of C and P in "C and possibly P" proceeds as follows. A tuple t belongs to the answer set of a query (1) if it satisfies the (crisp) condition given by

$$C(t) \textit{ and possibly } P(t) \equiv C(t) \wedge \exists s (C(s) \wedge P(s)) \Rightarrow P(t)$$

Notice that if *there is no conflict* between P and C *with respect to the content of a database*, i.e., there are tuples satisfying both of them, then the query collapses to $C \wedge P$, where $\wedge$ is the minimum (i.e. the conventional "and") while if there are no tuples satisfying both P and C, then the query collapses to C. Thus, clearly, in this interpretation of the bipolar query, the matching degree of a tuple t depends not only on t, but also on the whole set of tuples T which implies serious difficulties as the non-conventional "and possibly" aggregation cannot proceed via just the use of the values of the arguments.

The crucial issue is how to interpret the non-standard "and possibly" aggregation operator in the fuzzy context. We proposed three different ways to derive the logical formulas expressing the matching degree of a bipolar query with the "and possibly" operator (cf. Zadrożny and Kacprzyk [40]). Two of them rely on the treating the "and possibly" operator as a special case of the Chomicki's *winnow* operator [10, 40]. These three ways are the following

- by a direct fuzzification of (16):

$$C(t) \; \textit{and possibly} \; P(t) \equiv C(t) \wedge (\exists s \; (C(s) \wedge P(s)) \Rightarrow P(t)) \tag{16}$$

- by a direct fuzzification of the winnow operator (cf. [10]):

$$C(t) \; \textit{and possibly} \; P(t) \equiv C(t) \wedge \neg \exists s \; ((C(s) \wedge P(s) \wedge \neg P(t))) \tag{17}$$

- by using our fuzzy version of the winnow operator (cf. Zadrożny and Kacprzyk [40]):

$$C(t) \; \textit{and possibly} \; P(t) \equiv C(t) \wedge \forall s \; (C(s) \Rightarrow (\neg P(s) \vee P(t))) \tag{18}$$

In the classical Boolean logic all these three formulas are equivalent but this is not true in the context of fuzzy (multivalued) logic where different versions of the the conjunction and disjunction may be obtained using various t-norms and t-conorms, respectively [17]. We will consider the so-called De Morgan Triples $(\wedge, \vee, \neg)$ that comprise of a t-norm $\wedge$, a t-conorm $\vee$ and a negation $\neg$, where $\neg(x \vee y) = \neg x \wedge \neg y$ holds. The following three De Morgan Triples are of a special importance in fuzzy logic [17]: $(\wedge_{min}, \vee_{max}, \neg)$, $(\wedge_{\Pi}, \vee_{\Pi}, \neg)$, $(\wedge_W, \vee_W, \neg)$, called in what follows, respectively, the MinMax, Π and W triples, and the particular t-norms and t-conorms are

$t - norms$	
$x \;\wedge_{min}\; y = \min(x, y)$	*minimum*
$x \;\wedge_{\Pi}\; y = x \cdot y$	*product*
$x \;\wedge_{W}\; y = \max(0, x + y - 1)$	*Łukasiewicz*

$t - conorms$	
$x \;\vee_{max}\; y = \max(x, y)$	*maximum*
$x \;\vee_{\Pi}\; y = x + y - x \cdot y$	*probabilistic sum*
$x \;\vee_{W}\; y = \min(1, x + y)$	*Łukasiewicz*

and $\neg$ in all the above De Morgan Triples is $\neg x = 1 - x$.

In fuzzy logic, the universal and existential quantifier are meant, for the finite universes, to correspond to the maximum and minimum operators, respectively.

One can consider two implication operators related to the given De Morgan Triple $(\wedge, \vee, \neg)$, the so-called S-implications

$$x \rightarrow_{S-\vee} y = \neg x \vee y \tag{19}$$

and R-implications

$$x \rightarrow_{R-\wedge} y = \sup\{z : x \wedge z \leq y\} \tag{20}$$

Thus, for the particular De Morgan Triples, we obtain the following R-implication operators

$$\begin{aligned}
x \ \rightarrow_{R-min} \ y &= \begin{cases} 1 \text{ for } x \leq y \\ y \text{ for } x > y \end{cases} \\
x \ \rightarrow_{R-\Pi} \ y &= \begin{cases} 1 & \text{for } x = 0 \\ \min\{1, \frac{y}{x}\} & \text{for } x \neq 0 \end{cases} \\
x \ \rightarrow_{R-W} \ y &= \min(1 - x + y, 1)
\end{aligned}$$

and the following S-implication operators

$$\begin{aligned}
x \ \rightarrow_{S-max} \ y &= \max(1 - x, y) \\
x \ \rightarrow_{S-\Pi} \ y &= 1 - x + x \cdot y \\
x \ \rightarrow_{S-W} \ y &= \min(1 - x + y, 1)
\end{aligned}$$

Basically, in a series of our papers (cf. Zadrożny and Kacprzyk [39]), which culminated in Zadrożny and Kacprzyk [40], we analyzed in detail many aspects related to the choice of one of the formulas (16)–(18) to represent the bipolar queries and an appropriate modeling of the logical connectives occurring therein, i.e., the choice of one of the De Morgan Triples.

These are just some examples of more relevant properties that may be useful for the modeling of bipolar queries. For more information, we refer the reader to our recent paper [40].

Now, we will proceed to the very topic of this paper, namely the presentation of a new concept of compound bipolar queries combining traditional bipolar queries with queries with fuzzy linguistic quantifiers.

4 Compound Bipolar Queries: A Synergistic Combination of Bipolar Queries and Queries with Fuzzy Linguistic Quantifiers

In the discussion of the very essence and main properties of bipolar queries we have assumed up to now a simple, general form of a bipolar query, denoted by (C, P), meant as "C and possibly P". As we have already mentioned, the particular C's and P's can take on a more sophisticated form, namely can themselves be fuzzy queries with linguistics quantifiers. This gives rise to a new concept of a *compound bipolar query* proposed here as outlined in the Introduction.

Since the paper is meant to propose this new concept, and due to a lack of space, we will only present in a simple and illustrative way some more relevant formulations presented for the new compound bipolar queries. For clarity, we will strongly relate our discussion to the motivating example of real estate, and this will certainly not limit the generality.

For convenience of the reader, and clarity of presentation, let us repeat now the real estate example mentioned in the Introduction. Moreover, we will only use the approach based on a direct fuzzification of (16), i.e. of the source Lacroix and Lavency logical formulation of bipolar queries.

Our source bipolar query is

find a house which is *affordable* **and possibly** *well located*

which is written as [cf. due to (1)]

$$C \text{ and possibly } P \tag{21}$$

and, due to (16), we have

$$C(t) \textit{ and possibly } P(t) \equiv C(t) \wedge (\exists s\ (C(s) \wedge P(s)) \Rightarrow P(t))$$

The two types of conditions that exist in (21) are

- a required condition C, exemplified in the real estate case considered by "affordable",
- a desired condition P, exemplified in the real estate case by "well located".

We assume that the required and desired conditions involve a linguistic quantifier driven aggregation of some partial conditions (which directly correspond to attributes in a real estate database in question!), that is, correspond themselves to queries with fuzzy linguistic quantifiers.

For instance, in our real estate context

- the required condition "affordable" may be defined as (22), that is

$$Q \text{ of conditions among } \{c_i\}_{i=1,\ldots,n_C} \text{ are to be satisfied.} \tag{22}$$

where n_C is the number of conditions in C, exemplified by

Most of conditions among 'price IS *inexpensive*, bank loan IS *easy to get*, bank interest IS *not much higher than X%*, ...' are to be satisfied

- the desired condition "well located" may be defined as follows (cf. (5))

$$Q \text{ of conditions among } \{p_i\}_{i=1,\ldots,n_P} \text{ are to be satisfied.} \tag{23}$$

where n_P is the number of conditions in P, exemplified by

Most of conditions among 'distance to railroad station IS *short*, distance to bus stop IS *much less than 5 blocks*, number of buses at stop IS *high*, ...' are to be satisfied

The truth values of (22) and (23) are calculated due to—obviously assuming a regular nondecreasing quantifier (8)—as

- for the required condition C

$$\text{Truth}(Qic_i(t)) = \mu_Q(\frac{1}{n_C}\sum_{i=1}^{n_C}\mu_{c_i(t)}) \tag{24}$$

where n_C is the cardinality of the set of required conditions,
- for the desired condition P

$$\text{Truth}(Qip_i(t)) = \mu_Q(\frac{1}{n_P}\sum_{i=1}^{n_P}\mu_{p_i(t)}P(x_i)) \tag{25}$$

where n_P is the cardinality of the set of desired conditions.

Now, as we have mentioned, we will only use the approach based on a direct fuzzification of the source Lacroix and Lavency logical formulation of bipolar queries, i.e. (16).

Our query (C, P) can therefore be written as [cf. (1)]

$$C \text{ and possibly } P \tag{26}$$

which, since C and P are queries (conditions) with fuzzy linguistic quantifiers, i.e. $C(t) \equiv Q_1\ c_i(t)$ and $P(t) \equiv Q_2\ p_i(t)$, respectively, can further be written for simplicity as

$$Q_1c_i \text{ and possibly } Q_2p_i \tag{27}$$

which in turn implies due to (16)

$$Q_1\ c_i(t)\ \textit{and possibly}\ Q_2 p_i(t) \equiv Q_1 c_i(t) \wedge ((\exists s\ Q_1 c_i(s) \wedge Q_2 p_i(s)) \Rightarrow Q_2 p_i(t)) \tag{28}$$

We have limited our analysis, for simplicity, to C and P constituting queries with fuzzy linguistic quantifiers that are represented by linguistically quantified propositions without importance. One can clearly consider the case with importance, i.e. (11) with the truth value of the linguistically quantified proposition with importance calculated due to (12). Then, we only need to replace in (28) the respective truth values of P and C by those calculated due to (12).

5 Concluding Remarks

We have further elaborated on the concept of a compound bipolar query which we introduced conceptually in Kacprzyk and Zadrożny [22]. This new query type combines the concept of a bipolar query proposed by Dubois and Prade [13], followed by a fuzzy bipolar query due to Zadrożny [36] (cf. Zadrożny and Kacprzyk [40]), involving negative and positive information, notably meant as required and desired conditions with the concept of a query with fuzzy linguistic quantifiers proposed by Kacprzyk and Ziółkowski [24], and Kacprzyk et al. [25]. We show a new quality that can be gained by such a combination, and illustrate our arguments and approach on an easily comprehensible real estate example.

Acknowledgments This work was partially supported by the National Centre for Research (NCN) under Grant No. UMO-2012/05/B/ST6/03068.

References

1. Benferhat S, Dubois D, Kaci S, Prade H (2002) Bipolar representation and fusion of preferences on the possibilistic logic framework. In: Proceedings of the 8th international conference on principles and knowledge representation and reasoning (KR-02), Morgan Kaufmann, pp 421–448
2. Benferhat S, Dubois D, Kaci S, Prade H (2002) Bipolar possibilistic representations. In: Proceedings of the 18th conference in uncertainty in artificial intelligence, Morgan Kaufmann, pp 45–52
3. Benferhat S, Dubois D, Kaci S, Prade H (2006) Bipolar possibility theory in preference modeling: representation, fusion and optimal solutions. Inf Fusion 7:135–150
4. Benferhat S, Dubois D, Kaci S, Prade H (2008) Modeling positive and negative information in possibility theory. Int J Intell Syst 23:1094–1118
5. Bordogna G, Pasi G (1995) Linguistic aggregation operators of selection criteria in fuzzy information retrieval. Int J Intell Syst 10(2):233–248
6. Bosc P, Pivert O (1995) SQLf: a relational database language for fuzzy querying. IEEE Trans Fuzzy Syst 3(1):1–17

7. Bosc P, Pivert O (1992) Discriminated answers and databases: fuzzy sets as a unifying expression means. In: Proceedings of the IEEE international conference on fuzzy systems, pp 745–752
8. Bosc P, Pivert O (1993) An approach for a hierarchical aggregation of fuzzy predicates. In: Proceedings 2nd IEEE international conference on fuzzy systems, pp 1231–1236
9. Bosc P, Pivert O, Mokhtari A, Lietard L (2010) Extending relational algebra to handle bipolarity. In: Proceedings of the 2010 ACM symposium on applied computing (SAC), ACM, pp 1718–1722
10. Chomicki J (2002) Querying with intrinsic preferences. LNCS 2287:34–51
11. De Tré G, Zadrożny S, Bronselaer A (2010) Handling bipolarity in elementary queries to possibilistic databases. IEEE Trans Fuzzy Sets 18(3):599–612
12. De Tré G, Zadrożny S, Matthe T, Kacprzyk J, Bronselaer A (2009) Dealing with positive and negative query criteria in fuzzy database querying. LNCS 5822:593–604
13. Dubois D, Prade H (2002) Bipolarity in flexible querying. LNCS 2522:174–182
14. Dubois D, Prade H Handling bipolar queries in fuzzy information processing, vol. 18, pp 97–114
15. Dubois D, Prade H (2008) An introduction to bipolar representations of information and preference. Int J Intell Syst 23:866–877
16. Dubois D, Prade H (2009) An overview of the asymmetric bipolar representation of positive and negative information in possibility theory. Fuzzy Sets and Syst 160:1355–1366
17. Fodor J, Roubens M (1994) Fuzzy preference modelling and multicriteria decision support. Kluwer Academic Publishers, Dordrecht
18. Galindo J (ed) (2008) Handbook of research on fuzzy information processing in databases. Information Science Reference, New York
19. Grabisch M, Greco S, Pirlot M (2008) Bipolar and bivariate models in multicriteria decision analysis: descriptive and constructive approaches. Int J Intell Syst 23:930–969
20. Kacprzyk J, Zadrożny S (2001) Computing with words in intelligent database querying: standalone and internet-based applications. Inf Sci 134(1–4):71–109
21. Kacprzyk J, Zadrożny S (1995) FQUERY for access: fuzzy querying for a windows-based DBMS. In: Bosc P, Kacprzyk J (eds) Fuzziness in database management systems. Physica-Verlag, Heidelberg, pp 415–433
22. Kacprzyk J, Zadrożny S (2013) Hierarchical bipolar fuzzy queries: towards more human consistent flexible queries, professor of FUZ-IEEE'2013, Hyderabad, India, IEEE Explore, pp 1–8
23. Kacprzyk J, Zadrożny S (2013) Compound bipolar queries: combining bipolar queries and queries with fuzzy linguistic quantifiers. In: Proceedings of EUSFLAT-2013, Atlantis Press, pp 848–855
24. Kacprzyk J, Ziółkowski (1986) A Database queries with fuzzy linguistic quantifiers. IEEE Trans Syst Man Cybern 16:474–479
25. Kacprzyk J, Ziółkowski A, Zadrożny S (1989) FQuery III+: a "human consistent" database querying system based on fuzzy logic with linguistic quantifiers. Inf Syst 6:443–453
26. Lacroix M, Lavency P (1987) Preferences: putting more knowledge into queries. In: Proceedings of the 13 international conference on very large databases, Brighton, pp 217–225
27. Lietard L, Rocacher D, Bosc P (2009) On the extension of SQL to fuzzy bipolar conditions. In Proceedings of the NAFIPS-2009, pp 1–6
28. Lindström P (1966) First-order predicate logic with generalized quantifiers. Theoria 32:186–195
29. Mostowski A (1957) On a generalization of quantifiers. Fundam Math 44:12–36
30. Yager RR (1988) On ordered weighted averaging operators in multi-criteria decision making. IEEE Trans Syst Man Cybern 18:183–190
31. Yager R (1992) Higher structures in multi-criteria decision making. Int J Man-Mach Stud 36:553–570
32. Yager R (1996) Fuzzy logic in the formulation of decision functions from linguistic specifications. Kybernetes 25(4):119–130

33. Yager RR, Kacprzyk J (1997) The ordered weighted averaging operators: theory and applications. Kluwer, Norwell
34. Yorke M Bipolarity or not? Some conceptual problems related to bipolar rating. Br Educ Res J 27(2):171–186
35. Zadeh LA (1983) A computational approach to fuzzy quantifiers in natural languages. Comput Math Appl 9:149–184
36. Zadrożny S (2005) Bipolar queries revisited. In: LNAI 3558, pp 387–398
37. Zadrożny S, De Tré G, Kacprzyk J (2010) Remarks on various aspects of bipolarity in database querying. In: Proceedings of the DEXA'10, international workshops, IEEE computer society, pp 323–327
38. Zadrożny S, Kacprzyk J (2006) Bipolar queries and queries with preferences. In: Proceedings of the DEXA'06, pp 415–419
39. Zadrożny S, Kacprzyk J (2007) Bipolar queries using various interpretations of logical connectives. In: LNCS 4529, pp 181–190
40. Zadrożny S, Kacprzyk J (2012) Bipolar queries: an aggregation operator focused perspective. Fuzzy Sets Syst 196:69–81

Process Inspection by Attributes Using Predicted Data

Olgierd Hryniewicz

Abstract SPC procedures for process inspection by attributes are usually designed under the assumption of directly observed quality data. However, in many practical cases this direct observations are very costly or even hardly possible. For example, in the production of pharmaceuticals costly and time consuming chemical analyses are required for the determination of product's quality even in the simplest case when we have to decide whether the inspected product conforms to quality requirements. The situation is even more difficult when quality tests are destructive and long-lasting, as it is in the case of reliability testing. In such situations we try to predict the actual quality of inspected items using the values of predictors whose values are easily measured. In the paper we consider a situation when traditional prediction models based on the assumption of the multivariate normal distribution cannot be applied. Instead, for prediction purposes we propose to use some techniques known from data mining. In this study, which has an introductory character, and is mainly based on the results of computer simulations, we show how the usage of popular data mining techniques, such as binary regression, linear discrimination analysis, and decision trees may influence the results of process inspection.

1 Introduction

The majority of popular Statistical Process Control (SPC) tools have been designed under the assumption that the observed quality characteristics are independent, and usually normally distributed. Only few SPC procedures, such as the T^2 control chart, introduced by Hotelling in the 1947, were designed for the statistical inspection of processes described by multivariate data, see [6]. However, in modern production processes many process characteristics can be measured simultaneously, and very often the assumption of the multivariate normality of their observed values simply does not hold. Recently, in the works like [8, 12–14] some new techniques have been

O. Hryniewicz (✉)
Systems Research Institute, Newelska 6, 01-447 Warsaw, Poland
e-mail: hryniewi@ibspan.waw.pl

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_7

proposed for dealing with interdependent statistical quality data. They are mainly based on the so called profile data. In this approach a regression-type model that describes the relationship between the quality characteristic of interest and some other explanatory variables is built. Then, control charts are used for the control of stability of the parameters (profiles) of such regression models. These methods can be used for the analysis of different dependencies of a regression type, both linear and non-linear. However, in practically all cases the proposed models have been obtained under the assumption of the normality of measured characteristics. Moreover, it is assumed that all important quality characteristics of interest are *directly* measurable.

Automation of contemporary production processes allows for measurements of important parameters of produced items. When specification limits are set for the values of these measurements one can say that 100 % quality inspection has been implemented for this process. However, in many cases actual quality characteristics of interest cannot be measured during a process. For example, in the production of pharmaceuticals costly and time consuming chemical analyses are required for the determination of product's quality even in the simplest case when we have to decide whether the inspected product conforms to quality requirements. In such and similar cases 100 % inspection is hardly feasible. The same is when the measurement procedure is disruptive as in the case of the measurement of breaking strength. The most difficult, and practically impossible to implement, are measurements which are both disruptive and long-lasting. For example, [5] considered the case when the most important quality characteristic is the lifetime of a produced object. In all these cases a direct inspection of quality characteristics of interest is practically impossible. One can think, however, about an indirect inspection of such characteristics. In such a case we should measure other easily measurable characteristics, and use obtained values for the prediction of the unobserved values of the quality characteristic of interest.

The problem of an indirect inspection of important quality characteristics attracted attention of relatively few authors for the last more than fifty years. According to the most popular approach, introduced in a series of papers by Owen and collaborators, see [9], a multivariate probability distribution of the random vector $(Z, X_1, \ldots, X_k)$ is built, where Z is the quality characteristic of interest, and $X_1, \ldots, X_k$ are the characteristics which are directly measurable in a production process. In all these papers it was assumed that the vector $(Z, X_1, \ldots, X_k)$ is described by the multivariate (usually bivariate) normal (Gaussian) distribution describing $(Z, X_1, \ldots, X_k)$. Another approach is based on the assumption that the relation between the predicted random variable Z and the explanatory variables $X_1, \ldots, X_k$ is described by a certain (usually linear) regression model. Also in this case the normality assumption about Z is usually used in practice. In both cases there is a direct link of the proposed methods to the multivariate SPC tools mentioned in the first paragraph of this section.

Unfortunately, the models mentioned above are of rather limited applicability when the actual multivariate probability distribution of $(Z, X_1, \ldots, X_k)$ is different from the normal (Gaussian) one, and when the number of predictors (explanatory variables) $X_1, \ldots, X_k$ is not small. In order to cope with this problem [5] has proposed to use the data mining methodology. In his proposal the simplest case when the random variable Z is described by the Bernoulli distribution is considered. In such a case a classifier (e.g. linear classifier, decision tree or artificial neural network) is built using training data consisted of the limited number of observations $(Z, X_1, \ldots, X_k)$, and this classifier is used in the inspection process for "labeling" the produced items. In statistical quality control this type of inspection is named "by attributes". One can think, however, about the application of this methodology for a more general setting when the variable Z is real-valued. In such a case one can use such data mining methods as e.g. regression trees for the prediction of its unobserved values.

Classifiers used in the inspection process are usually built using small amount of data, named training data. Thus, the results of classification are not error-free. What is more important, however, that the relation between the results of classification and the actual level of the quality characteristic of interest may be quite complicated. Therefore, there is a need to investigate the impact the quality of the classification procedures on the efficiency of SPC procedures used in production processes. This problem was first approached in [5]. In this paper we extend his results and consider a more general model of the process inspection.

The remaining part of this paper is organized as follows. In Sect. 2 we describe an assumed mathematical model of the inspection process and discuss the problem of the prediction of directly unobserved quality characteristics using data mining techniques. In Sect. 3 we describe a simulation model used for the evaluation of three prediction algorithms, namely the Binary Regression, the Linear Discrimination Analysis (LDA) and the Classification Decision Tree Quinlan's C4.5 algorithm. In Sect. 4 the performance of these algorithms is evaluated in terms of prediction errors for both non-shifted and shifted process levels. Then, in Sect. 5, we compare the efficiency of the considered prediction algorithms for two types of 100 % process inspection methods. Some conclusions and indication for a future work are presented in the last section of the paper.

2 Mathematical Model of a Process with Predicted Values of Quality Characteristics

A general mathematical model of a process with predicted values of quality characteristics is a simple one. Let $Z_1, \ldots, Z_p$ be p quality characteristics whose values are not directly observed using measurement tools available for process operators. Their values should be predicted using observations $X_1, \ldots, X_k$ of k predictors. It can be done if we assume their common probability distribution, i.e. the probability distribution of a vector $(Z_1, \ldots, Z_p, X_1, \ldots, X_k)$. According to the famous Sklar's theorem this distribution is univocally described by a $(p + k)$-dimensional copula, and marginal probability distributions of $Z_1, \ldots, Z_p$ and $X_1, \ldots, X_k$. Unfortunately, this general model is hardly applicable in practice. Therefore, we need to propose a model that is simpler and more easy for practical interpretation. One of such possible simple models was proposed in [5]. This model has a hierarchical 3-level structure.

On the top level there is a one-dimensional real-value quality characteristic T. However, for purpose of quality inspection we are not interested in the values of T, but in the values of a binary (attribute) variable defined as

$$Z_t = \begin{cases} 0\ , \ T \geq t \\ 1\ , \ T < t \end{cases} \tag{1}$$

This model has a direct interpretation when items produced in the monitored process are considered as either conforming or nonconforming to certain quality requirements. In his paper [5] considers a process where the life-time of produced items is the actual quality characteristic of interest. However, we can never identify its precise probability distribution. Instead, we can identify the probability distribution of a random variable T describing the results of an accelerated life test (ALT) or even a highly accelerated life-time test (HALT). Basing on previously accumulated experience we assume that items whose predicted life-time T which would be observed in the ALT test is smaller than a certain value t are potentially less reliable, and can be considered as nonconforming. Thus, one can believe that the inference about the values of Z_t is more robust to uncertainties related to the unknown relationship between the actual life-time and the predicted value of the life-time that would be observed in the accelerated life-time tests.

The first level of our model describes the predictors $X_1, \ldots, X_k$. In order to make it simpler for future simulations we assume that consecutive pairs of predictors (X_i, X_{i+1}), $i = 1, \ldots, k-1$ are described by $k-1$ copulas $C_i(F_i(X_i), F_{i+1}(X_{i+1}))$, $i = 1, \ldots, k-1$, where $F_1(X_1), \ldots, F_k(X_k)$ are the cumulative probability functions of the marginal distribution of the predictors. In our model we have to assume the type of the proposed copulas, and the strength of dependence between the pairs of random variables whose joint two-dimensional probability distributions are described by these copulas. Pearson's coefficient of correlation r is often used for this purpose. However, its applicability is limited to the case of the classical multivariate normal distribution, or—in certain circumstances—to the case of the multivariate elliptic distributions (for more information see [2]). When dependent random variables cannot be described by such a model, and it is a usual case when the predicted variable is the life-time, a non-parametric measures of dependence should be used. In our model we propose to use Kendall's coefficient of association τ defined, in its population version in terms of copulas, as (see [7])

$$\tau(X, Y) = 4 \int \int_{[0,1]^2} C(u, v) dC(u, v) - 1. \tag{2}$$

Numerical comparisons of the values of Pearson's r, Kendall's τ, and Spearman's ρ presented in [4] show that the usage of Pearson's r in the analysis of reliability data may lead to wrong conclusions, and Kendall's τ is in this case a much better measure of dependence.

In order to have a more realistic model for simulation purposes [5] proposed to use an in-between second level of latent variables $HX_1, \ldots, HX_k$. Each hidden variable HX_i is associated with the predictor variable X_i and its fictitious realizations are measured at the same scale as the predicted random variable T. The dependence between HX_i and X_i is described by the copula $C_{Hi}(F_{Hi}(HX_i), F_i(X_i))$. Moreover, we assume that there exists a certain linear relationship between the expected value of HX_i and the expected value of X_i. This assumption is needed if we want to model the effects of the shifts in the expected values of the predictors on the expected value of the predicted variable T which is related to the hidden variables by a certain, possibly nonlinear, function

$$T = f(HX_1, \ldots, HX_k). \tag{3}$$

The probability distribution of T, and hence the probability distribution of Z_t, can be observed only in specially designed experiments. In practice, however, we can observe only their predicted values T' and Z_t', respectively. As in production processes we are mainly interested in the binary classification of produced items for prediction purposes we may use several, say s, classifiers, $K_1, \ldots, K_s$, each of the form

$$Z_t' = K(X_1, \ldots, X_k) \tag{4}$$

In practical applications one can choose only one appropriate classifier or an ensamble of classifiers with a predetermined decision rule (e.g. majority voting).

3 Process Inspection

Statistical methods for process inspection have been successfully used in industrial practice since the works of Shewhart in 1920s. They usually have a form of control charts that allow to visualize the results of measurements of random samples taken from the inspected process. In contemporary production processes, however, all produced may be inspected automatically, and the observed values of measured quality characteristics have a form of respective time series.

The main aim of statistical process control (SPC) is to keep the process under control. From a mathematical point of view it means that the probability distributions of the quality characteristics of interest should be stable in time. In practice it is required that their expected values and variances should be constant in time. Shifts in the expected values (in one or both directions) and the increase of variance are considered as the signs of the process deterioration, and appropriate correction actions should be taken in order to make the process stable.

From a statistical point of view SPC processes can be viewed as the procedures for the detection of a change-point of a process. Since the introduction of sequential statistical methods (late 1940s and early 1950s) many statistical procedures related to the change-point problem have been proposed, both in classical and Bayesian setting. A good overview of basic methods can be found in the book [1], available on the internet. The most useful methods have been implemented in commercial software packages, and in recently developed R package *changepoint* available at the repository CRAN. Recently, the problem of the change-point detection has also attracted attention of the Artificial Intelligence community (e.g., see [15]).

Despite the abundance of good methods for the change-point detection practitioners in industry prefer to use old and very simple methods such as, e.g., control charts. Even if a production process is 100 % inspected the results of inspection can be analyzed and visualized using a well-known control charts. The CUSUM control chart seems to be the most appropriate chart for this purpose. However, for some practitioners it looks as too complicated and not easy for interpretation. Therefore, in this study we consider two simpler approaches used in practice.

In the first approach the process is divided into consecutive segments of the constant length of n elements. In the second approach only the last n produced items are taken into account, so the inspection process uses a kind of a sliding window of the width n. The results of such inspection are analyzed on a (usually virtual) control chart. In the first case it is a classical Shewhart control chart. In the second case it is a Moving Average (MAV) chart with a Shewhart-like control limits. In both cases an alarm signal is generated if the observed value of the fraction nonconforming falls beyond the control limits calculated from the data taken from a process which is in a stable state.

In the case of the inspection by attributes the control limits of the Shewhart chart are calculated from a simple formula

$$CL_{Sh} = \hat{p} \pm 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{5}$$

where $\hat{p}$ is the fraction of nonconforming items, estimated from a process data. In the case of the MAV chart the consecutive values displayed on the chart are strongly correlated. Therefore, for the construction of the MAV control chart with a Shewhart-type (i.e., three sigma) control limits we have to estimate not only the value of $\hat{p}$, but the value of its standard deviation $\sigma_{\hat{p}}$ as well. If we estimate these values the control limits of the MAV chart are given by

$$CL_{MAV} = \hat{p} \pm 3\sigma_{\hat{p}}. \tag{6}$$

One has to note, that the construction of the Shewhart control chart is easier from a practical point of view. When we calculate the value of $\hat{p}$ we do not need to retain the original data, and we can use this value for the design of control charts for segments of a different size. In the case of the MAV control chart we have to recalculate from the original data the value of $\sigma_{\hat{p}}$ if we want to change the width of a sliding window.

4 Process Simulation

The general model described in Sect. 3 does not allow, due to its complexity, to infer about important characteristics of the inspection process. A large computer simulation program that can be used for the Monte Carlo evaluation of many characteristics which are important to practitioners has been developed. In the current version the program implements the model described in Sects. 3 and 4 for only four predictors. Moreover, this implementation is somewhat restricted as some of its important features are typical for the case when reliability characteristics of produced items have to be predicted in the inspection process.

The simulation program consists of three modules. First module generates a stream of data points, i.e. the values of predictors, the value of the unobserved output variable, and the class associated with this value. The second module builds classification (prediction) algorithms, and computes the predicted class of the output variable using the values of predictors as its input. The third module simulates the process of inspection, i.e. building and operating a control chart.

In the first module of the simulation program the probability distributions of predictors defined by a user on the first level of the model can be chosen from a set of five distributions: uniform, normal, exponential, Weibull, and log-normal. Each of these distributions is represented either as a two-parameter parameter single distribution or as a mixture of two such distributions. This second option enables a user to model more complex distributions such as, e.g., bimodal ones. For the second level of the model a user can choose the probability distributions of the hidden variables from a set of distributions which are typically used for the description of reliability data (i.e., defined on the positive part of the real line): exponential, Weibull, and log-normal. The dependence between the pairs of predictors, and between predictors and associated hidden variables can be described by the following copulas: independent, normal, Clayton, Gumbel, Frank, and Fairlie-Gumbel-Morgenstern (FGM). The detailed description of these copulas can be found, e.g., in [7]. The strength of this dependence is defined by the value of Kendall's coefficient of association τ. The expected values of the distributions of the hidden variables in this simulation model depend in a linear way on the expected values of its related predictors. Finally, on the third level, hidden random variables are transformed to the final random variable T. The relation between the hidden variables and T is strongly non-linear, and is described by operators of a "min-max" type.

Several types of classifiers (several linear regression models, two Linear Discriminant Analysis procedures, and one decision tree classifier) have been implemented in the second module of the simulation program. The classifiers are built using samples of size n_t of training data consisted of the vectors of the values of predictors (x_1, x_2, x_3, x_4), and the actual value of the assigned class. In this paper we consider only three of them which represent three different general approaches to the classification problem.

The first considered classifier is a simple binary linear regression. We label the considered classes by 0 and 1, respectively, and consider these labels as real numbers, treating them as observations of a real dependent variable in the linear regression model of the following form:

$$R = w_1 * X_1 + w_2 * X_2 + w_3 * X_3 + w_4 * X_4 + w_0, \tag{7}$$

where R is the predicted class of an item described by explanatory variables X_1, X_2, X_3, X_4, and w_1, w_2, w_3, w_4, w_0 are respective coefficients of regression equation estimated from a training set of n_t elements. The value of R estimated from (7) is a real number, so an additional requirement is needed for the final classification (e.g. if $R < 0,5$ an item is classified as belonging to the class 0, and to the class 1 otherwise). The only advantage of this naive method is its simplicity. In the problem of binary classification the logistic regression is definitely a better choice, but the classical linear regression is implemented in all spreadsheets, such as, e.g., MS Excel. For this reason we have chosen this classifier as the easiest to implement in practice.

The next classifier implements the algorithm of the Linear Discriminant Analysis (LDA) introduced by Fisher, and described in many textbooks on multivariate statistical analysis and data mining (see, e.g. [3]). In this method statistical data are projected on a certain hyperplane estimated from the training data. New data points which are closer to the mean value of the projected on this hyperplane training data representing the class 0 than to the mean value of training data representing the remaining class 1 are classified to the class 0. Otherwise, they are classified to the class 1. The equation of the hyperplane is given by the following formula:

$$L = y_1 * X_1 + y_2 * X_2 + y_3 * X_3 + y_4 * X_4 + y_0, \tag{8}$$

where L is the value of the transformed data point calculated using the values of the explanatory variables X_1, X_2, X_3, X_4, and y_1, y_2, y_3, y_4, y_0 are respective coefficients of the LDA equation estimated from a training set of n_t elements. If Z_L denote the decision point, a new item is classified to the class 0 if $L \leq Z_L$, and to the class 1 otherwise. In our simulation we implemented the classical method of the calculation of Z_L, and this point is just the average of the mean values of the transformed data points from the training set that belonged to the class 0 and the class 1, respectively. The calculation of the LDA equation (8) is not so simple. However, it can be done using basic versions of many statistical packages such as SPSS, STATISTICA, etc. Therefore, we have implemented this classifier in order to represent methods available in basic versions of professional statistical packages of general purpose.

The third considered classifier is based on the implementation of the one of the most popular data mining classification algorithms, namely the classification decision tree (CDT) algorithm C4.5 introduced by [10], and described in many textbooks on data mining, such as, e.g., [11]. In our simulations we used its version (known as J48) implemented in the WEKA software, available from the University of Waikato, Hamilton, New Zealand, under the GNU license. The decision tree is constructed using "IF..THEN..ELSE" rules, deducted from the training data. In this paper for the description of the CDT classifier we use the notation of the MS Excel function

$IF(lt, t, f)$, where lt is a logical condition (e.g. $C < 50$), t is the action when $lt = true$, and f is the action when $lt = false$. The actions t and f can be defined as the combinations of other *IF* functions, or—finally—as the assignments of classes to the considered items.

The third module of the program is dedicated to the simulation of a production process. The simulated process is described by two mathematical models. First model describes the process in the so called Phase I when it is considered as being under control. Second model represents a process that has been deteriorated, and is considered as not being under control. Each simulation run begins with the simulation of a sample of n_d items which are used for the estimation of the parameters of the inspection procedure (Shewhart control charts or MAV control charts). Separate control charts are designed for actual and predicted (using different classifiers) data. Next, $n - 1$ items (we call them historical data) are simulated using a first model in order to represent the process before its possible transition to the possible out of control state. This additional sample is needed if we want to simulate the MAV control chart. After the historical data have been generated we start to simulate consecutive items using either the first model (for the analysis of process' behavior when it is under control) or the second model (for the analysis of process' behavior when it is out of control). The first item generated in this way forms, together with the historical data, the first sample whose fraction of nonconforming items (actual and predicted using different classifiers) is compared with the control lines of the respective control charts. If the value of one (or more) of these fractions falls beyond the control limit of the respective control chart an alarm signal is generated. Consecutive items are generated until the moment when the alarm signals have been observed on all considered control charts. Note, that for the MAV chart decisions are taken after the generation of each consecutive item. In contrast to the MAV chart, for the Shewhart chart decisions are taken after the generation of consecutive segments of n items.

In order to evaluate statistical properties of the simulated inspection processes we have to repeat the simulation process N_R times. In the context of 100 % inspection the most important characteristic of the inspection process is its *Time to Signal* (TS). By the Time to Signal we understand the number of inspected items between the moment of process deterioration and the alarm signal. The data obtained from N_R simulation runs let us to analyze the probability distributions of *TS*, and to compute the estimated values of important characteristics such as its average value (ATS), median, standard deviation, and skewness.

The simulation program described above can be run using different settings of parameters. In the paper by [5] only one particular setting was used. We have decided to continue this work using the same setting of parameters. This allows us not to repeat some interesting findings that have been already described in [5]. In the model used in the simulation experiments described in this paper the random predictor X_1 is distributed according to the normal distribution, X_2 has the exponential distribution, X_3 is distributed according to the log-normal distribution, and X_4 has the Weibull distribution. The dependence between X_1 and X_2 has been described by the Clayton copula with $\tau = 0, 8$. The joint distribution of X_2 and X_3 is described by the normal copula with $\tau = -0, 8$ (Notice that this is bivariate "normal" distribution, but with non-normal marginals!), and the joint distribution of X_3 and X_4 has been described by the Frank copula with $\tau = 0, 8$. The hidden variable HX_1 is described by the log-normal distribution, and its joint probability distribution with X_1 has been described by the normal copula with $\tau = -0, 8$. The joint distribution of HX_2 and X_2 has been described by the Frank copula with $\tau = 0, 9$, and the marginal distribution of HX_2 is assumed to be the exponential. The joint model of HX_3 and X_3 is similar, but the copula describing the dependence in this case is the Gumbel copula. Finally, the hidden variable H_D is described by the Weibull distribution, and its joint probability distribution with D has been described by the Clayton copula with $\tau = -0, 8$. The random variable T that describes the life time has been defined as $T = min[max(HX_1, HX_2), min(HX_3, HX_4)]$. The parameters of the aforementioned distributions have been found experimentally in such a way, that the items belonging to class 1 have the values of T smaller than 5, and to class 0 otherwise. Moreover, the relation between the predictors X_1, X_2, X_3, X_4 and their hidden counterparts HX_1, HX_2, HX_3, HX_4 is such that a shift in the expected value of each observed variable, measured in terms of its standard deviation, results with the similar shift of the expected value of its hidden counterpart, measured in terms of its own standard deviation.

Theoretically, classifiers can be built using training data for each production run. However, the collection of training data is usually very costly (it requires, e.g., making long-lasting destructive tests), and the same classifier can be used for several production runs. In our experiment we have decided to use ten sets of classifiers. Each of these sets consisted of classifiers designed using the same training data set of $n_t = 100$ elements. Note that in the data mining community this size of the training data is considered as small, and usually insufficient. However, in production reality this is often the upper limit for the number of elements which can be used for the experimental building of the prediction model.

The coefficients of the regression equation (7) are presented for the considered ten data sets in Table 1. Those coefficients that have been indicated by the regression statistical tool as statistically non-significant have been printed in this Table in *italics*. However, we have to remember that in the calculation of the significance of regression coefficient it is assumed that observed data are distributed according to the normal probability distribution. In our case it is obviously not true, so in our classification experiment we have used full regression equations.

Just a first look at Table 1 reveals that the estimated regression equation (7) may be completely different, depending on the chosen training data set. However, some general pattern is visible: only explanatory variable X_2 (represented by the coefficient w_2) is significant for all regression lines. On the other hand, the explanatory variable X_3 (represented by the coefficient w_3) seems to be of no practical importance in the classification process.

A similar comparison of the decision model parameters in the LDA case is presented in Table 2. In this case we cannot say about statistical significance of the parameters of the decision rule. However, the general impression is the same as in the case of the regression algorithm. The particular models look completely different depending on the training data set. However, in all the cases the explanatory variable

Table 1 Regression model—different sets of training data

Dataset	w_1	w_2	w_3	w_4	w_5
Set 1	*−0.133*	−0.034	*−0.0001*	−0.026	3.036
Set 2	*0.010*	−0.039	*0.0001*	−0.033	2.321
Set 3	*0.144*	−0.033	*<0.0001*	*0.0002*	*1.359*
Set 4	*−0.194*	−0.034	*−0.0005*	*−0.019*	3.396
Set 5	−0.346	−0.021	−0.0006	*0.002*	3.763
Set 6	*0.073*	−0.047	*<0.0001*	*−0.023*	2.058
Set 7	*−0.142*	−0.040	−0.0004	*−0.0008*	3.019
Set 8	*0.109*	−0.038	*<0.0001*	*−0.001*	1.611
Set 9	−0.346	−0.039	*0.0002*	−0.034	4.054
Set 10	*−0.002*	−0.018	−0.0005	0.042	1.745

Table 2 Linear discrimination analysis—different sets of training data

Dataset	y_1	y_2	y_3	y_4	y_0	Midpoint
Set 1	0.687	0.174	0.001	0.133	−6.338	0.628
Set 2	−0.045	0.178	−0.001	0.151	−2.710	−0.014
Set 3	−0.646	0.148	<0.0005	−0.001	1.663	0.464
Set 4	0.880	0.152	0.002	0.087	−7.499	0.585
Set 5	1.500	0.091	0.003	−0.008	−9.121	0.254
Set 6	−0.342	0.219	<0.0005	0.107	−1.399	0.706
Set 7	0.703	0.196	0.002	0.037	−6.044	0.784
Set 8	−0.501	0.173	<0.0005	0.006	0.636	0.629
Set 9	1.458	0.127	0.001	0.143	−10.048	0.344
Set 10	0.008	0.087	0.002	−0.206	0.272	0.771

X_3 seems to have no effect (very low values of the coefficient describing this variable) on the classification.

Finally, let us considered different decision rules estimated for the CDT algorithm. Because of a completely different structure of decision rules presented in Table 3 we cannot compare directly these rules with the rules described by Eqs. (7)–(8). They also look completely different for different training data sets, but in nearly all cases (except for the Set 9) decision are predominantly (and in two cases exclusively) based on the value of the explanatory variable X_3. One can notice that the weights assigned to the explanatory variables in the CDT algorithm are nearly exactly opposite to the weights assigned in the classification models (7)–(8). In order to explain this shocking difference one should take into account that both the linear regression and the LDA procedures are based on the assumption of the linear dependence and normality. Hryniewicz [5] shows the example of training data from which it is clear that in the considered model the dependence between the random variable T and the

Table 3 Decision trees—different sets of training data

Dataset	Decision rule
Set 1	*IF*(X_3 <= 70,0181; *IF*(X_3 <= 56,1124; 1; *IF*(X_3 <= 63,2962; 2; 1)); *IF*(X_4 <= 16,4381; 2; *IF*(X_1 <= 4,3509; 2; 1)))
Set 2	*IF*(X_3 <= 56,4865; 1; *IF*(X_4 <= 17,3301; 2; *IF*(X_1 <=4,0217; 2; 1)))
Set 3	*IF*(X_3 <= 73,6148; *IF*(X_3 <= 57,1355; 1; *IF*(X_1 <= 5,0876; 2; *IF*(X_4 <= 4,497; 2; 1))); *IF*(X_4 <= 17,3499; 2; 1))
Set 4	*IF*(X_3 <= 70,2191; 1; *IF*(X_4 <= 15,9098; 2; 1))
Set 5	*IF*(X_3 <=73,1584; 1; (*IF*(X_4 <= 17,0516; (*IF*(X_3 <= 87,8921; (*IF*(X_4 <= 5,0679; 2; 1)); 2)); 1)))
Set 6	*IF*(X_3 <= 60,3912; 1; *IF*(X_4 <= 16,3504; 2; 1))
Set 7	*IF*(X_3 <= 71,8184; 1; 2)
Set 8	*IF*(X_3 <= 71,4456; 1; (*IF*(X_3 <= 983,0929; 2; (*IF*(X_4 <= 18,8213; 2; 1)))))
Set 9	*IF*(X_2 <= 14,7339; (*IF*(X_4 <= 16,7482; (*IF*(X_4 <= 4,527; 1; 2)); 1)); 1)
Set 10	*IF*(X_3 <= 60,5044; 1; 2)

explanatory variables X_3 and X_4 is definitely not linear, but also non-monotonic. This dependence cannot be captured by the measures of linear correlation in the linear models (7)–(8). Moreover, this example shows that the explanatory potential of these two variables is seemingly much greater than the potential of the variables X_1 and X_2. We will discuss this problem in the next section of this paper.

In the experiment we used a different coding of the observed classes ((2,1) instead of (0,1)), and this feature is reflected in the values of coefficients presented Tables 1 and 2, and the description of rules presented in Table 3. However, this change does not influence the properties of the considered classifiers.

In our simulation experiment we have simulated N_R runs of the inspected process. Each of the ten sets of classifiers was used a random number of times $N_{R,j}, j = 1, \ldots, 10$ having the same expectation $N_R/10$. The joint probability distribution of $N_{R,1}, \ldots, N_{R,10}$ was multinomial with parameter N_R, and with all probabilities equal to $1/10$.

5 Simulation Experiments—Selected Results

5.1 General Remarks

The problem of process inspection using predicted binary attribute data can be decomposed into the following subproblems of great practical importance:

- Accuracy of classifiers in the presence of strongly non-linear dependence and non-normal measurement data;
- Influence of imperfect classification on observed process quality levels;
- Effectiveness of different inspection policies based on the predicted attributes.

Each of these problems should be subdivided into parts. First, when the process data are generated by the same model as the data used for the design of classifiers. Second, when these two types of data are generated by different mechanisms. The second of these problems is seldom considered by the data mining community where it is usually assumed that the model discovered from training data is the same as the model that describes future data. In the area of statistical process control (SPC) this assumption may not hold, as we may face processes (or parts of them) which, by definition, are described by models different from those discovered from the training data.

The problems mentioned above have been analyzed using the results of many simulation experiments. In the following subsections we will present only a part of these results which seems to us the most interesting. It is important to notice, however, that all of the presented here results have been obtained for *one* model of the process, described in Sect. 4. Therefore their interpretation should be rather of *qualitative* character. They should be considered rather as the signalization of important problems than already established solutions.

5.2 Accuracy of Classifiers

There exist many methods that can be used for the evaluation of the quality of classifiers. Some of them are simple and natural, some others (e.g. the methods based on the analysis of the ROC characteristic) are much more complicated, and more difficult to interpret. In this research we have decided to use five simple to calculate quality measures of classifiers: Accuracy, Sensitivity, Precision, Specificity, and the F1 index. Definitions of these measures can be found in every book on data mining, and in the Internet (Wikipedia). The results presented below have been obtained from 1000 simulation runs. In each simulation run a sample of 5000 items was generated, and evaluated by one set of classifiers taken randomly from 10 sets described in Sect. 4. Therefore, the results presented below represent averages taken on the population of ten different sets of classifiers.

In Tables 4, 5, 6, 7 and 8 we distinguish nine cases. The case labeled *Sh0* represents the situation when both the training set and the evaluation set are generated from the same model. In the SPC context it means that the inspected process is under control. The cases labeled *ShXim-05s* represent the situation when the evaluation set was generated by the the model for which the expected value of the predictor X_i has been shifted by $-0,5\sigma$, where σ is the standard deviation of X_i. Similarly, the cases labeled *ShXip-05s* represent the situation when the evaluation set was generated by the the model for which the expected value of the predictor X_i has been shifted by $+0,5\sigma$. In these tables RegBin stands for the binary regression classifier, LDA stands for the LDA classifier, and C4.5 stands for Quinlan's decision tree C4.5 classifier.

In the 45 comparisons (5 indices times 9 types of processes) the decision tree C4.5 classifier was the best 27 times. Moreover, for 9 types of processes it has dominated the remaining classifiers 7 times, and in neither of these types it was dominated by

Table 4 Average accuracy of implemented classifiers

Process	RegBin	LDA	C4.5
Sh0	0.861	0.858	**0.910**
ShX1m-05s	0.860	0.863	**0.910**
ShX1p-05s	0.858	0.846	**0.910**
ShX2m-05s	0.839	0.861	**0.911**
ShX2p-05s	0.869	0.842	**0.909**
ShX3m-05s	0.828	**0.853**	0.667
ShX3p-05s	**0.866**	0.844	0.861
ShX4m-05s	0.845	0.848	**0.892**
ShX4p-05s	**0.866**	0.846	0.834

Table 5 Average sensitivity of implemented classifiers

Process	RegBin	LDA	C4.5
Sh0	0.548	0.718	**0.830**
ShX1m-05s	0.520	0.686	**0.821**
ShX1p-05s	0.564	0.736	**0.838**
ShX2m-05s	0.407	0.595	**0.830**
ShX2p-05s	0.626	0.786	**0.832**
ShX3m-05s	0.468	0.647	**0.943**
ShX3p-05s	0.594	**0.764**	0.519
ShX4m-05s	0.500	0.653	**0.712**
ShX4p-05s	0.586	0.765	**0.838**

Table 6 Average precision of implemented classifiers

Process	RegBin	LDA	C4.5
Sh0	**0.859**	0.733	0.829
ShX1m-05s	**0.877**	0.760	0.836
ShX1p-05s	**0.844**	0.711	0.824
ShX2m-05s	**0.912**	0.818	0.829
ShX2p-05s	0.800	0.667	**0.828**
ShX3m-05s	**0.934**	0.846	0.495
ShX3p-05s	0.795	0.661	**0.857**
ShX4m-05s	0.878	0.764	**0.882**
ShX4p-05s	**0.832**	0.695	0.655

Table 7 Average specificity of implemented classifiers

Process	RegBin	LDA	C4.5
Sh0	0.967	0.905	**0.937**
ShX1m-05s	**0.975**	0.923	0.939
ShX1p-05s	**0.957**	0.883	0.935
ShX2m-05s	**0.985**	0.950	0.938
ShX2p-05s	**0.942**	0.861	0.935
ShX3m-05s	**0.984**	0.943	0.548
ShX3p-05s	0.949	0.869	**0.965**
ShX4m-05s	0.633	0.698	**0.781**
ShX4p-05s	**0.957**	0.871	0.832

Table 8 Average F1-index of implemented classifiers

Process	RegBin	LDA	C4.5
Sh0	0.664	0.719	**0.823**
ShX1m-05s	0.650	0.717	**0.820**
ShX1p-05s	0.664	0.710	**0.825**
ShX2m-05s	0.558	0.680	**0.824**
ShX2p-05s	0.715	0.717	**0.822**
ShX3m-05s	0.619	**0.727**	0.640
ShX3p-05s	0.673	**0.700**	0.624
ShX4m-05s	0.633	0.698	**0.781**
ShX4p-05s	0.678	0.714	**0.720**

any of its competitors. Therefore, for the assumed model of dependence it obviously outperforms its competitors. This result is not unexpected, as the assumed model of the process was strongly non-linear and non-normal. The performance of a simple binary regression classifier is also quite good, and this confirms the opinion presented in Sect. 4.2 of [3] that in the case of only two distinguished classes this classifier performs well. This may suggest the usage of the ensemble of three classifiers with the majority of voting decision rule. What is somewhat unexpected is poor efficiency of the LDA classifier whose behavior is usually similar (or even better—especially in the case of more than 2 considered classes) to that of based on linear regression. The possible explanation of this phenomenon can be related to non-normality of input data, as the calculation of the decision criterion in LDA strongly depends on this assumption (see [3]).

5.3 Process Quality Levels for Predicted Observations

The accuracy of the considered classifiers is, as it is seen from the results presented in Sect. 5.2, far from being perfect. Therefore, the observed process quality levels, expressed in terms of the fraction of nonconforming items, may be quite different from the actual one. In Table 9 we show the estimates of the fraction nonconforming p estimated from samples of 1000 elements. The entries of Table 9 represent the averages obtained over the population of the considered ten sets of classifiers in 1000 simulation runs. The column labeled *Actual* represents the values obtained if the actual values of the quality characteristic were directly observed.

From Table 9 one can immediately see that in many cases the observed levels of process' quality are different from the actual ones. This is due to erroneous results of classification where we can distinguish two cases: erroneous classification of nonconforming items as conforming (false positives), and erroneous classification of conforming items as nonconforming (false negatives). This problem was discussed in details in [5]. When the fraction of false positives is larger than the fraction of false negatives is larger, then the observed process' quality level will be smaller than the actual one, and vice versa. Classification errors of both types should be avoided, but their consequences could be different. When the observed process' quality level is smaller than the actual one, and it is always the case when the binary regression classifier is used, nonconforming items (e.g., potentially unreliable) pass quality inspection. When consequences of non-detecting nonconforming items are serious this situation should be avoided. On the other hand, if the observed process' quality level is larger than the actual one we face losses incurred by the false rejection of good items.

The comparison of observed and actual quality levels yields only partial information about the efficiency of inspection. Consider, for example the case when the percentages of false positives and false negatives are large but equal. In this case the observed and the actual quality levels will be the same, but economic consequences

Table 9 Observed values of the fraction nonconforming p

Process	Actual	RegBin	LDA	C4.5
Sh0	0.252	0.163	0.252	0.257
ShX1m-05s	0.253	0.151	0.232	0.253
ShX1p-05s	0.252	0.178	0.275	0.260
ShX2m-05s	0.252	0.114	0.188	0.256
ShX2p-05s	0.253	0.208	0.303	0.259
ShX3m-05s	0.303	0.153	0.236	0.601
ShX3p-05s	0.234	0.178	0.280	0.148
ShX4m-05s	0.270	0.155	0.170	0.229
ShX4p-05s	0.244	0.176	0.284	0.331

for the process' owner may be disastrous. Therefore, we should look for other features which could help us in distinguishing between "good" and "bad" classifiers. One has to note, however, than the terms "good" and "bad" are strongly context-dependent. Let us consider how the performance of classifiers influences the ability of the inspection process to detect process' disruptions. From Table 9 we see that shifts in the expected values of two predictors, X_1 and X_2, practically do not change the actual fraction nonconforming. However, the observed fractions of nonconforming items are changing quite significantly when RegBin and LDA classifiers are used for prediction purposes. The consequences of this situation can be detrimental, either due to the increasing rate of false alarms (when the observed fraction of nonconforming items increases) or the increasing fraction of accepted nonconforming items (when the observed fraction of nonconforming items decreases). Another dangerous situation we observe in the case of the shift in the expected value of the explanatory variable X_4. For all considered classifiers the observed fractions of nonconforming items are changing but in the exactly opposite direction that the change of the actual value of this quality characteristic.

A more complex situation is noticed when the shift in the expected value of the predictor X_3 is observed, as it can be seen from Fig. 1.

When this expected value decreases the actual fraction of nonconforming items significantly increases. This phenomenon cannot be detected when RegBin and LDA classifiers are used, but is even amplified when we use the C4.5 classifier for prediction purposes. Therefore, when this classifier is used, the deterioration of the process will be quickly detected. However, as "free lunches" do not exist, we have to pay in this case for the increased fraction of false positives when the actual fraction of nonconforming items is decreasing. A more detailed explanation of this problem can be found in [5].

While discussing the consequences of the usage of the C4.5 classifier one can think if the analysis of popular measures of the quality of classification may be

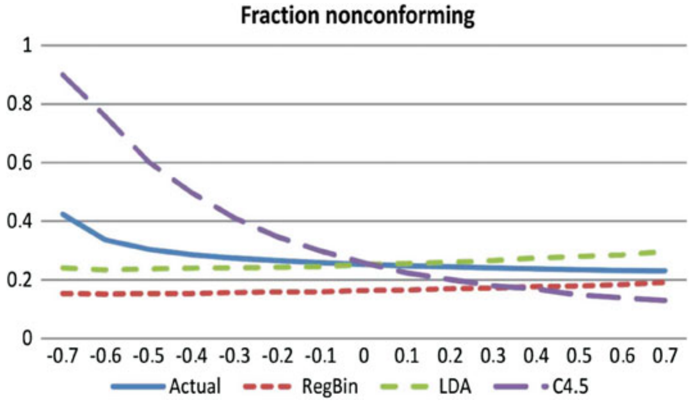


Fig. 1 Observed fraction nonconforming in the presence of shift in X_3 (multiples of σ)

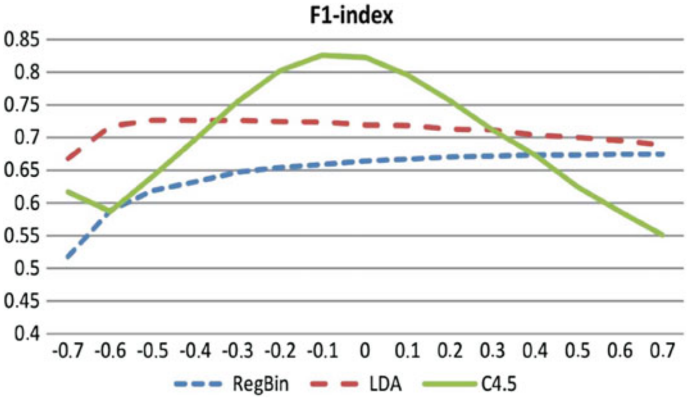


Fig. 2 Values of F1-index in the presence of shift in X_3 (multiples of σ)

useful. On Fig. 2 we see how the value of the popular F1-index changes with the change of the expected value of X_3, and thus with the change of the actual fraction of nonconforming items. We see that for processes being in "in-control" or "nearly in-control" state the decision tree C4.5 classifier visibly outperforms its competitors. However, in the presence of large (both positive and negative) shifts in the expected value of X_3 the situation is quite different. When we try to interpret the consequences of this "bad" performance we can find that in the case of negative shifts this erroneous behavior is even useful, and in the case of positive shifts the negative consequences are quite improbable (processes usually do not improve spontaneously their quality). Therefore, the analysis of the measures of the quality of classification should be accompanied by the analysis of the consequences of incorrect classification.

5.4 Properties of Inspection Policies

The main goal of the research described in this paper is to develop an efficient inspection procedure for a production process when the quality of produced items is observed indirectly. When classical SPC tools, such as control charts, are used there are two main questions to be answered. First, how to design the chosen control chart. Second, what is the ability of this chart to detect process' deterioration. In the problem considered in this paper we have to add a third question about the influence of the quality of used classifiers on the effectiveness of process' inspection.

In the classical SPC the effectiveness of a control chart is measured by its Average Run Length (ARL) defined as the average number of samples taken between the occurrence of deterioration and the alarm or as the average number of samples taken between consecutive alarms when the process is under control. When we use SPC procedures for the analysis of the results of 100 % inspection we have to use the

Average Time to Signal (ATS) characteristic. ATS in this case is defined as the average number of items produced between the occurrence of deterioration and the alarm or as the average number of items produced between consecutive alarms when the process is under control.

When quality of produced items is evaluated by attributes (either conforming or nonconforming) Shewhart p-charts with one-sided (upper) control limits are usually used. Charts with two-sided control limits are used only when we want to detect improvements of the process (e.g. when the input raw material is changed to a new and supposedly better one). In our experiments we have considered both types of control charts. However, for reasons explained in the previous section (actual deterioration of a process can be signaled as its "improvement"—see the case of the X_4 predictor) we have decided to suggest the usage of two-sided control charts.

Let us consider the application of the Shewhart control chart for the process inspection. First, let us consider the case when the inspected process is under control. In Table 10 we present the values of ATS (usually denoted $ATS0$) for different sizes of samples (segments of the process). The parameters of the control chart have been calculated using the estimated value of the fraction of nonconforming items estimated from the sample of 1000 elements taken from a stable process.In the column labeled "Actual" we present the values of ATS of the chart for the actual, but not observed, values of the quality characteristic.

The similar values for the MAV control chart are presented in Table 11.

The results presented in Tables 10 and 11 are strikingly different, but this difference is not difficult to explain. In the case of the Shewhart control chart decisions are taken after observing a sample of n elements. Thus, the value of, e.g., $ATS0 = 30000$ when the sample size is, e.g., equal to 100 means that on average 300 samples are evaluated before the alarm (This is the value of ARL!). In the case of the MAV control chart the decision is taken after each produced item. Thus, the value of ATS is the same as the value of ARL. It means that the "waiting time" in terms of the number of taken decisions is in the case of the MAV chart much larger than in the case of the Shewhart chart, but the relationship between the respective values of the average time to signal (ATS) is just opposite.

Note that in the case of a stable inspected process each alarm is a false one. Therefore, we should prefer larger sample sizes n in order to have these alarms not so frequently. This can be achieved by the widening of distance between control

Table 10 Average time to signal $ATS0$—Shewhart chart

Sample size	Actual	RegBin	LDA	C4.5
100	31119	34055	32883	31526
200	48194	50710	50338	49556
300	65289	73189	71630	64340
400	75432	77902	78753	70359
500	93990	92629	92638	86232

Table 11 Average time to signal *ATS*0—MAV chart

Sample size	Actual	RegBin	LDA	C4.5
100	4158	4700	4091	4201
200	6593	7317	6314	6742
300	8715	9596	8431	8486
400	10828	11004	10650	10037
500	12221	11513	12521	11761

Table 12 Average time to signal *ATS* (shifted process)—Shewhart chart

Sample size	Actual	RegBin	LDA	C4.5
100	3359	39669	27618	2935
200	3204	50710	37381	4592
300	4008	63318	40791	5699
400	3140	60572	43864	6781
500	2835	68963	45039	11445

Table 13 Average time to signal *ATS* (shifted process)—MAV chart

Sample size	Actual	RegBin	LDA	C4.5
100	638	5247	3837	501
200	815	6349	4879	562
300	782	8746	5382	729
400	905	9864	6430	657
500	744	10234	6284	2030

limits on a chart. The effect of such a change may be, unfortunately, detrimental if we want to detect the deterioration of the inspected process as quickly as possible. In Tables 12 and 13 we present the values of *ATS* when the expected value of the explanatory variable (predictor) X_3 is shifted downwards by $0,5\sigma$. From the second column of the Table 9 we see that this shift results in the increase of the fraction nonconforming by 20 %. This a really severe deterioration of the process and should be detectes as quickly as possible.

From the analysis of simulation results presented in Tables 10, 11, 12 and 13 we see that the inspection is effective only in the case when the decision tree classifier C4.5 is used for the prediction of quality of inspected items. When the LDA classifier is used the inspection process allows to detect deterioration but with visibly smaller efficiency. The binary regression RegBin classifier is in the considered case completely ineffective. From the analysis of Tables 4, 5, 6, 7 and 8 we see that in the considered case the decision tree C4.5 classifier in comparison to its competitors is characterized by a larger value of Sensitivity and smaller values of Precision and Specificity. The same is when we compare the LDA and the RegBin classifiers. This behavior of classifiers can be explained by noting that high sensitivity and low precision and specificity describe the situation when the observed percentage of nonconforming items is larger than the actual one. Therefore, in the case of process deterioration the probability of alarm increases, and the value of *ATS* decreases. One should note, however, that in the case of a stable (under control) process the observed value of the fraction of nonconforming items is also larger than the actual one. This phenomenon results with a somewhat misleading information about the actual process level, but does not inflict the probability of false alarm (and the *ATS*0 value), as the control limits are designed on the basis on observed but not actual values of the fraction of nonconforming items.

All the results described in this paper represent averages calculated with respect to different sets of classifiers. From more detailed results, presented in [5] for the case of the inspection based on the Shewhart *p*-chart, one can find that depending on the instance of the training set alarms may be triggered when the actual impact of shifts in explanatory variables on actual quality is negligible, and—vice versa—may not be triggered when it is needed. This behavior strongly depends upon the type of a classifier, and its parameters estimated from a training data. Moreover, In this paper we assumed that alarms are triggered by crossing either the lower or the upper control limit. When only the upper control limit of the control chart is active, the respective values of the ATS are much larger, especially in the case of no-shift or when the shift in the explanatory variable has a small effect on the quality variable of interest. The situation is even worse when the deterioration of the process is accompanied with lowering of the observed fraction of nonconforming items, as it is the case in the upwards shift of the explanatory variable X_4. In such a case such deterioration may be never noticed using considered statistical methods.

6 Conclusions

The results presented in this paper add important information to that already given in [5]. However, this information is still of a very preliminary character, as the results from simulation experiments represent only one particular model of a process. They confirm the findings presented in [5] that in the case of non-normal distributions of quality characteristics, non-linear dependencies between observable (explanatory), and not directly observable (only predicted!) values quality characteristics of processes the inspection procedures based on control charts may be not effective. The most popular classifiers that are used for prediction purposes may not perform well, and their performance is difficult to be predicted in advance. Further research is needed with the aim to find ensambles of classifiers that can be more effective than single classifiers in finding process' deterioration. Such ensambles have to be robust to the change of the model of data used in their design.

References

1. Basseville M, Nikiforov IV (1993) Detection of abrupt changes: theory and applications. Prentice-Hall, Englewood Cliffs
2. Embrechts P, Lindskog F, McNeil A (2003) Modelling dependence with copulas and applications to risk management. In: Rachev S (ed) Handbook of heavy tailed distributions in finance, Chapter 8. Elsevier, Amsterdam, pp 329–384
3. Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning. data mining, inference, and prediction, 2nd edn. Springer, New York
4. Hryniewicz O, Karpiński J (2014) Prediction of reliability—pitfalls of using Pearson's correlation. Eksploatacja i Niezawodnosc—Maintenance Reliab 16:472–483
5. Hryniewicz O (2015) SPC of processes with predicted data—application of the data mining methodology, In: Knoth S, Schmid W (eds) Frontiers in statistical quality control—12, Physica Verlag, Heidelberg, pp 219–235
6. Montgomery DC (2011) Introduction to statistical quality control, 6th edn. Wiley, New York
7. Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York
8. Noorsana R, Saghaei A, Amiri A (2011) Statistical analysis of profile monitoring. Wiley, Hoboken
9. Owen DN, Su YH (1977) Screening based on normal variables. Technometrics 19:65–68
10. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, Los Altos
11. Witten IH, Frank E, Hall MA (2011) Data mining. practical machine learning tools and techniques, 3rd edn. Elsevier, Amsterdam
12. Woodall WH, Spitzner DJ, Montgomery DC, Gupta S (2004) Using control charts to monitor process and product profiles. J Qual Techn 36:309–320
13. Wang YT, Huwang L (2012) On the monitoring of simple linear berkson profiles. Qual Rel Engin Int 28:949–965
14. Xu L, Wang S, Peng Y et al (2012) The monitoring of linear profiles with a GLR control chart. J Qual Techn 44:348–362
15. Yamada M, Kimura A, Naya F, Sawada H (2013) Change-point detection with feature selection in high-dimensional time-series data. In: Proceedings of the 23rd International Joint. Conference on Artificial Intelligence, Beijing, pp 1827–1833

Székely Regularization for Uplift Modeling

Szymon Jaroszewicz and Łukasz Zaniewicz

Abstract Uplift modeling is a subfield of machine learning concerned with predicting the causal effect of an action at the level of individuals. This is achieved by using two training sets: treatment, containing objects which have been subjected to an action and control, containing objects on which the action has not been performed. An uplift model then predicts the difference between conditional success probabilities in both groups. Uplift modeling is best applied to training sets obtained from randomized controlled trials, but such experiments are not always possible, in which case treatment assignment is often biased. In this paper we present a modification of Uplift Support Vector Machines which makes them less sensitive to such a bias. This is achieved by including in the model formulation an additional term which penalizes models which score treatment and control groups differently. We call the technique Székely regularization since it is based on the energy distance proposed by Székely and Rizzo. Optimization algorithm based on stochastic gradient descent techniques has also been developed. We demonstrate experimentally that the proposed regularization term does indeed produce uplift models which are less sensitive to biased treatment assignment.

1 Introduction

The aim of conventional classification methods is to predict the class membership probabilities of new objects' based on a given training dataset. In practice, however, usually a more important question is how this probability changes as a result of some action. Modeling this particular change (or difference) is the scope of *uplift modeling*. In contrast to traditional response modeling, uplift approach uses two training sets:

S. Jaroszewicz (✉) · Ł. Zaniewicz
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
e-mail: s.jaroszewicz@ipipan.waw.pl

Ł. Zaniewicz
e-mail: l.zaniewicz@phd.ipipan.waw.pl

S. Jaroszewicz
National Institute of Telecommunications, Warsaw, Poland

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_8

treatment dataset with data on objects on which a particular action was taken, and control with data on untreated cases.

Probably the most intuitive example of utility of uplift modeling is a direct marketing campaign for a certain product. It is easy to see that we can divide customers into four groups:

1. Customers who purchased the product *because* they were targeted and would not have purchased otherwise.
2. Customers who would have purchased the product regardless of whether they were targeted or not.
3. Customers who would not have purchased the product regardless of whether they were targeted or not.
4. Customers who were going to purchase the product, but were annoyed by the action and, as a result, changed their mind.

In the first group the action is clearly beneficial, and in the fourth clearly detrimental. In the second and third groups the action has no real effect and those two groups can in fact be aggregated into a single neutral group. Similar problems arise in personalized medicine where a treatment may bring recovery from a disease, but the patient may also be exposed to dangerous side effects.

In contrast, traditional classification models are only able to distinguish two groups: those who respond (*after*, not necessarily *because* of the action) and those who do not. In many real world situations, this distinction does not correspond to the problem actually being solved.

1.1 Biased Treatment Assignment Problem

A very important aspect of uplift modeling is how the cases are assigned to treatment and control groups. The best scenario is a randomized controlled experiment, where the assignment is random and does not depend on the features of the cases. Unfortunately, such an experiment is not always possible (e.g. for ethical or financial reasons) or only historical data may be available where, for example, the treatment was applied to patients which a doctor considered most suitable.

If treatment assignment was not random and biased then the effect of the action cannot, usually, be estimated directly. Consider, for example, a medical treatment with potentially serious side effects. The doctor might then decide not to apply it to patients in severe condition who will thus be placed in the control group. However, such cases are also more likely not to recover from the disease making the control group outcomes look worse and the treatment more effective than it is in reality.

In this paper we present Uplift Support Vector Machines originally proposed in [22] with an additional penalty term, which we call *Székely regularizer*. As a result, we obtain uplift models which are additionally forced to make similar predictions on the treatment and control groups, thus helping to reduce the effect of treatment assignment bias.

The additional regularizer is based on so called energy distance between probability distributions proposed by Székely and Rizzo [6, 18, 19]. The distance has the property that it is zero if and only if the distributions are identical, it can thus be used to enforce similar distributions of model scores in the treatment and control groups.

2 Related Work

Uplift modeling has received relatively little attention in the literature. The first paper mentioning it explicitly was [10] where decision trees designed specifically for that problem were discussed. More details on the method were later presented in [11].

A trivial approach to uplift modeling uses two probabilistic classifiers, one built on the treatment dataset, the other on control, whose predicted probabilities are then subtracted. The approach may however suffer from a serious drawback: both models will focus on predicting class probabilities in both groups and ignore the (usually much weaker) differences between them. A good example can be found in [11]. Therefore, most research in uplift modeling has been concerned with approaches which model the conditional difference in success probabilities directly.

Many such approaches are based on adaptations of decision trees. For example, in [11] uplift trees have been proposed which are based on a statistical test of success rate differences after the split and in [15, 16] trees based on information theoretical divergences between treatment and control class probabilities. Ensembles of decision trees have been described in [2]; a more thorough analysis of various types of ensembles in the uplift setting can be found in [17].

Regression based techniques can be found for example in [4], where a class variable transformation is presented which allows for converting uplift modeling problems into classification problems. Similar techniques have been discussed in the statistical literature [12, 21].

In [22] Uplift Support Vector Machines have been proposed which allow for explicit identification of cases for which the action is positive, neutral and negative. The model is described in the next section. Another type of uplift SVMs was proposed in [8] and is based on direct maximization of the area under the uplift curve.

Good overviews of uplift modeling can be found in [11] and in [16]. Procedures for correcting treatment assignment bias will be discussed in Sect. 5.3.

2.1 Uplift Support Vector Machines

Let us first introduce some notation. Vectors will be denoted with boldface, lowercase letters, $\mathbf{x}$, $\mathbf{w}$. A dataset is a collection of records $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i$ is the ith feature vector and $y_i \in \{-1, 1\}$ the class value for the ith record. The outcome 1 is considered the successful or desired outcome. The superscript ^{T} will be used to denote terms related to the treatment group and the superscript ^{C} terms related to the control. For example, the treatment training dataset is $\mathbf{D}^T = \{(\mathbf{x}_i^T, y_i^T) : i = 1, \ldots, n^T\}$ and the control training set is $\mathbf{D}^C = \{(\mathbf{x}_i^C, y_i^C) : i = 1, \ldots, n^C\}$.

We now discuss in more details the Uplift Support Vector Machines presented in [22] which we will use as the starting point for further developments. The machine is based on two separating hyperplanes

$$H_1 : \langle \mathbf{w}, \mathbf{x} \rangle - b_1 = 0, \qquad H_2 : \langle \mathbf{w}, \mathbf{x} \rangle - b_2 = 0.$$

The model predictions are made according to the following formula:

$$M(\mathbf{x}) = \begin{cases} +1 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle > b_1 \text{ and } \langle \mathbf{w}, \mathbf{x} \rangle > b_2, \\ 0 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle \leq b_1 \text{ and } \langle \mathbf{w}, \mathbf{x} \rangle > b_2, \\ -1 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle \leq b_1 \text{ and } \langle \mathbf{w}, \mathbf{x} \rangle \leq b_2, \end{cases} \tag{1}$$

that is, the model classifies the effect of the action on a point $\mathbf{x}$ as positive $(+1)$, negative (-1), or neutral (0). A graphical interpreation of the model is shown in Fig. 1 (taken from [22]). The hyperplane H_1 separates positive and neutral predictions and the hyperplane H_2 separates neutral and negative predictions.

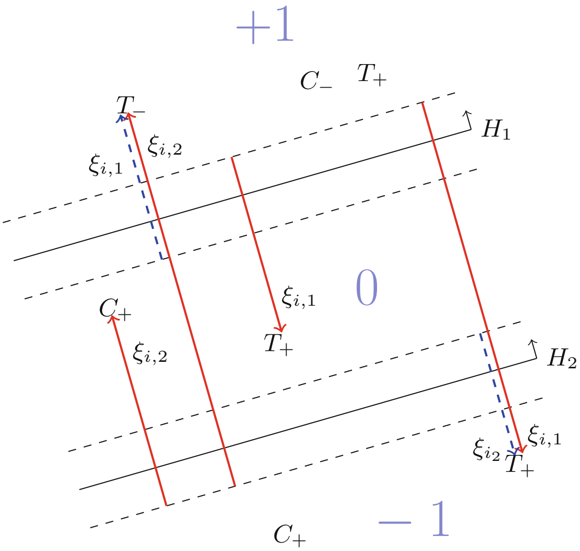


Fig. 1 The Uplift SVM optimization problem. Example points belonging to the positive class in the treatment and control groups are marked respectively with T_+ and C_+. Analogous notation is used for points in the negative class. The figure shows penalties incurred by points with respect to the two hyperplanes of the USVM. Positive sides of hyperplanes are indicated by small arrows at the right ends of lines in the image. *Red solid arrows* denote the penalties incurred by points which lie on the wrong side of a single hyperplane, *blue dashed arrows* denote additional penalties for being misclassified also by the second hyperplane

Let us now formulate the optimization task which allows for finding the model's parameters $\mathbf{w}, b_1, b_2$. We will use $\mathbf{D}_+^T = \{(\mathbf{x}_i, y_i) \in \mathbf{D}^T : y_i = +1\}$ to denote data points belonging to the positive class in the treatment group and $\mathbf{D}_-^T = \{(\mathbf{x}_i, y_i) \in \mathbf{D}^T : y_i = -1\}$ to denote points in that group belonging to the negative class. Analogous notation is used for points in the control group.[1]

The version presented here is slightly different than that given in [22]: the soft margin penalties are averaged separately over the treatment and control groups. As a result both groups have the same impact on the optimized risk. The optimization problem is to find weights $\mathbf{w}$ maximizing the function $R(\mathbf{w})$ defined as

$$\begin{aligned} R(\mathbf{w}) = \frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle &+ \frac{C_1}{n^T}\sum_{\mathbf{D}_+^T}\xi_{i,1} + \frac{C_2}{n^T}\sum_{\mathbf{D}_-^T}\xi_{i,1} + \frac{C_2}{n^T}\sum_{\mathbf{D}_+^T}\xi_{i,2} + \frac{C_1}{n^T}\sum_{\mathbf{D}_-^T}\xi_{i,2} \\ &+ \frac{C_1}{n^C}\sum_{\mathbf{D}_-^C}\xi_{i,1} + \frac{C_2}{n^C}\sum_{\mathbf{D}_+^C}\xi_{i,1} + \frac{C_2}{n^C}\sum_{\mathbf{D}_-^C}\xi_{i,2} + \frac{C_1}{n^C}\sum_{\mathbf{D}_+^C}\xi_{i,2}, \end{aligned} \tag{2}$$

subject to constraints

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x}_i\rangle - b_1 &\geq +1 - \xi_{i,1}, \text{ for all } (\mathbf{x}_i, y_i) \in \mathbf{D}_+^T \cup \mathbf{D}_-^C, \\ \langle \mathbf{w}, \mathbf{x}_i\rangle - b_1 &\leq -1 + \xi_{i,1}, \text{ for all } (\mathbf{x}_i, y_i) \in \mathbf{D}_-^T \cup \mathbf{D}_+^C, \\ \langle \mathbf{w}, \mathbf{x}_i\rangle - b_2 &\geq +1 - \xi_{i,2}, \text{ for all } (\mathbf{x}_i, y_i) \in \mathbf{D}_+^T \cup \mathbf{D}_-^C, \\ \langle \mathbf{w}, \mathbf{x}_i\rangle - b_2 &\leq -1 + \xi_{i,2}, \text{ for all } (\mathbf{x}_i, y_i) \in \mathbf{D}_-^T \cup \mathbf{D}_+^C, \\ \xi_{i,j} &\geq 0, \text{ for all } i = 1, \ldots, n,\ \ j \in \{1, 2\}. \end{aligned}$$

Note that the model has two penalty coefficients, C_1 and C_2. The properties of the model are given in detail in [22], here we only review the main results without proofs, which easily carry over to the modified formulation given in this paper. First, the model is valid iff $b_1 \geq b_2$, this is the case when $C_2 \geq C_1$ which puts a constraint on the values of C_1 and C_2. The role of the coefficient C_1 is the same as in classical SVMs. From Fig. 1 it is clear that the coefficient C_2 determines the additional penalty for points which are on the wrong side of both hyperplanes (e.g. a treatment point with negative outcome which is classified as positive). It turns out that the ratio C_2/C_1 determines the proportion of neutral predictions. For $C_1 = C_2$ no points are classified as neutral and for a sufficiently large value of C_2/C_1 almost all points are.

In the following sections we will add an additional penalty term to (2) which will force similar model behavior in the treatment and control groups.

[1] The values of the class variable should not be confused with model predictions defined in (1). For example, a model prediction of +1 means that we expect the class variable to take the value of +1 if the action is performed ($y^T = +1$) and to take the value of −1 if the action is not performed ($y^C = -1$).

3 Székely Regularized Support Vector Machines

One way to view an uplift model is as a function which maps feature vectors into the set $\{-1, 0, +1\}$. The value is interpreted as a decision on whether the action applied to a given case will be beneficial, neutral, or detrimental. Another approach is for the model to return a *score*: a real number being an increasing function of the predicted probability that the action will be beneficial. In this paper we are going to define our Uplift Support Vector Machines using the discrete prediction model but for testing and regularization purposes we will use the linear score $\langle \mathbf{w}, \mathbf{x} \rangle$.

3.1 Distribution of Scores in Controlled Randomized Experiments

Let us now state an important property of score based uplift models used in controlled randomized experiments. Let M_s be an uplift model returning a score and $M_s(\mathbf{x})$ the score returned by the model for a specific instance $\mathbf{x}$. When the feature vector $\mathbf{x}$ is picked at random from a population distribution, then $M_s(\mathbf{x})$ is a random variable. Suppose $\mathbf{x}^T$ is picked at random from the treatment population and $\mathbf{x}^C$ from the control population. In a randomized controlled trial $\mathbf{x}^T$ and $\mathbf{x}^C$ follow the same distributions and therefore *$M_s(\mathbf{x}^T)$ and $M_s(\mathbf{x}^C)$ are random variables following the same distributions.* If the treatment assignment is not random, the distributions of $\mathbf{x}^T$ and $\mathbf{x}^C$ differ and so may those of $M_s(\mathbf{x}^T)$ and $M_s(\mathbf{x}^C)$.

In this paper we will use this property to obtain models which are less sensitive to treatment assignment bias. This will be achieved by adding a regularization term penalizing models which yield different score distributions in the treatment and control training sets.

3.2 The Energy Distance

In this paper we make use the concept of *energy distance* (also called *E-statistic*) $e^{(\alpha)}$ proposed in 2005 by Székely and Rizzo [6, 18–20]. Initially this concept was introduced as a measure of distance between clusters, but it is in fact a general a statistical distance between two or more probability distributions. The name comes from the fact that it was first derived from physics; later Székely applied this concept to statistics.

Let $A = \{\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_{n_1}\}$, $B = \{\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_{n_2}\}$ be two nonempty sets of points in $\mathbb{R}^d$. Formally, $e^{(\alpha)}(A, B)$ is defined as

$$e^{(\alpha)}(A,B) = \frac{n_1 n_2}{n_1+n_2}\left[\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|\mathbf{a}_i-\mathbf{b}_j\|^{\alpha} - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|\mathbf{a}_i-\mathbf{a}_j\|^{\alpha} - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|\mathbf{b}_i-\mathbf{b}_j\|^{\alpha}\right], \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm and α is parameter that influences the behavior of the distance. If $\alpha = 2$ then the distance is equal to zero iff the means of A and B are equal. The case $\alpha \in (0, 2)$ is more interesting, since the distance is then equal to zero iff the sets A and B are equal. Moreover, if A and B are random samples and $\alpha \in (0, 2)$, then, as the size of A and B grows to infinity, the distance between them tends to zero iff A and B are drawn from the same distribution (for $\alpha = 2$ the distributions from which they are drawn only need to have equal means). This property is important for the task the distance will be used for in this paper. Notice also that for $d = 1$ the Euclidean norms reduce to absolute values.

3.3 Model Formulation

We modify the risk function of Uplift Support Vector Machines by adding an extra term responsible for penalizing the difference in score distributions in the treatment and control groups. We call this term the *Székely regularization term*. The risk function of the regularized USVMs is

$$\begin{aligned}R(\mathbf{w}) = \frac{1}{2}\langle\mathbf{w},\mathbf{w}\rangle &+ \frac{C_1}{n^T}\sum_{\mathbf{D}_+^T}\xi_{i,1} + \frac{C_2}{n^T}\sum_{\mathbf{D}_-^T}\xi_{i,1} + \frac{C_2}{n^T}\sum_{\mathbf{D}_+^T}\xi_{i,2} + \frac{C_1}{n^T}\sum_{\mathbf{D}_-^T}\xi_{i,2}\\ &+ \frac{C_1}{n^C}\sum_{\mathbf{D}_-^C}\xi_{i,1} + \frac{C_2}{n^C}\sum_{\mathbf{D}_+^C}\xi_{i,1} + \frac{C_2}{n^C}\sum_{\mathbf{D}_-^C}\xi_{i,2} + \frac{C_1}{n^C}\sum_{\mathbf{D}_+^C}\xi_{i,2}\\ &+ C_3 S(\mathbf{D}^T,\mathbf{D}^C,\mathbf{w}).\end{aligned} \tag{4}$$

The risk is optimized subject to the same constraints as the risk given in (2). Above, $S(\mathbf{D}^T, \mathbf{D}^C, \mathbf{w})$ is the *Székely regularizer* given by

$$\begin{aligned}S(\mathbf{D}^T,\mathbf{D}^C,\mathbf{w}) =& \frac{2}{n^T n^C}\sum_{i=1}^{n^T}\sum_{j=1}^{n^C}|\langle\mathbf{w},\mathbf{x}_i^T\rangle - \langle\mathbf{w},\mathbf{x}_j^C\rangle|^{\alpha}\\ &- \frac{1}{(n^T)^2}\sum_{i=1}^{n^T}\sum_{j=1}^{n^T}|\langle\mathbf{w},\mathbf{x}_i^T\rangle - \langle\mathbf{w},\mathbf{x}_j^T\rangle|^{\alpha}\end{aligned}$$

$$- \frac{1}{(n^C)^2} \sum_{i=1}^{n^C} \sum_{j=1}^{n^C} |\langle \mathbf{w}, \mathbf{x}_i^C \rangle - \langle \mathbf{w}, \mathbf{x}_j^C \rangle|^\alpha. \tag{5}$$

Note that $\langle \mathbf{w}, \mathbf{x}_i \rangle$ is the score assigned by the model to a data record $\mathbf{x}_i$, and (5) is thus the energy distance (3) applied to the sets of scores assigned by the model to records in the treatment and control groups. Due to the properties of the energy distance the term will penalize models for which distributions of scores in both groups differ. The factor C_3 determines the strength of the penalty. The fraction $\frac{n^T n^C}{n^T + n^C}$ from (3) is absorbed into C_3 for the ease of exposition.

Let us now discuss the choice of α. Since we want to guarantee equal score distributions we need $\alpha \in (0, 2)$ [19]. However for $\alpha < 1$ the function S exhibits strong non-convexity and is thus more difficult to optimize. We should, therefore, choose α from the interval $[1, 2)$. We found values close to 1 to work better in practice but for $\alpha = 1$ the function S is not differentiable. We thus settled for $\alpha = 1.1$ which gives good properties and a smoother function to optimize. Note, however, that S may not be convex even for $\alpha \in [1, 2)$.

4 Optimization

We now describe the method used to optimize (4) subject to the constraints specified below (2). As a first step we rewrite the problem as an unconstrained optimization problem using the hinge loss:

$$\begin{aligned} R(\mathbf{w}) =& \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle \\ &+ \frac{C_1}{n^T} \sum_{\mathbf{D}_+^T} h(y_i^T(\langle \mathbf{w}, \mathbf{x}_i^T \rangle - b_1)) + \frac{C_2}{n^T} \sum_{\mathbf{D}_+^T} h(y_i^T(\langle \mathbf{w}, \mathbf{x}_i^T \rangle - b_2)) \\ &+ \frac{C_2}{n^T} \sum_{\mathbf{D}_-^T} h(y_i^T(\langle \mathbf{w}, \mathbf{x}_i^T \rangle - b_1)) + \frac{C_1}{n^T} \sum_{\mathbf{D}_-^T} h(y_i^T(\langle \mathbf{w}, \mathbf{x}_i^T \rangle - b_2)) \\ &+ \frac{C_2}{n^C} \sum_{\mathbf{D}_+^C} h(-y_i^C(\langle \mathbf{w}, \mathbf{x}_i^C \rangle - b_1)) + \frac{C_1}{n^C} \sum_{\mathbf{D}_+^C} h(-y_i^C(\langle \mathbf{w}, \mathbf{x}_i^C \rangle - b_2)) \\ &+ \frac{C_1}{n^C} \sum_{\mathbf{D}_-^C} h(-y_i^C(\langle \mathbf{w}, \mathbf{x}_i^C \rangle - b_1)) + \frac{C_2}{n^C} \sum_{\mathbf{D}_-^C} h(-y_i^C(\langle \mathbf{w}, \mathbf{x}_i^C \rangle - b_2)) \\ &+ C_3 S(\mathbf{D}^T, \mathbf{D}^C, \mathbf{w}), \end{aligned} \tag{6}$$

where h is the *hinge loss* function given by

$$h(q) = \max\{0, 1 - q\}.$$

We are going to use the Averaged Stochastic Gradient Descent algorithm [9] in order to optimize (4). The reason is that the algorithm is fast, stable and works well with non-smooth functions. Note that in our optimization problem the derivatives of the target function are not guaranteed to exist so methods such as conjugate gradient descent are not applicable.

In order to optimize the expression given in (6) we first need to compute its subgradient (since $R(\mathbf{w})$ is not everywhere differentiable there are values of $\mathbf{w}$ for which the gradient does not exist). Note that the subgradient of the hinge loss $h(q)$ is

$$\frac{\partial h(q)}{\partial q} = \begin{cases} -1 & \text{if } q < 1, \\ \text{any value in } [-1, 0] & \text{if } q = 1, \\ 0 & \text{if } q > 1. \end{cases}$$

Since for $q = 1$ any value in $[-1, 0]$ can be picked we will simply set

$$\frac{\partial h(q)}{\partial q} = -\mathbb{1}_{[1-q>0]}. \tag{7}$$

We can now give the expression for the subgradient of the minimized risk given in (6)

$$\begin{aligned}
&\frac{\partial R(\mathbf{w})}{\partial \mathbf{w}} = \\
&\quad \mathbf{w} - \frac{C_1}{n^T} \sum_{\mathbf{D}^T_+} \mathbb{1}_{[1-y_i^T(\langle \mathbf{w}, \mathbf{x}_i^T \rangle - b_1)>0]} \mathbf{x}_i^T y_i^T - \frac{C_2}{n^T} \sum_{\mathbf{D}^T_+} \mathbb{1}_{[1-y_i^T(\langle \mathbf{w}, \mathbf{x}_i^T \rangle - b_2)>0]} \mathbf{x}_i^T y_i^T \\
&\quad - \frac{C_2}{n^T} \sum_{\mathbf{D}^T_-} \mathbb{1}_{[1-y_i^T(\langle \mathbf{w}, \mathbf{x}_i^T \rangle - b_1)>0]} \mathbf{x}_i^T y_i^T - \frac{C_1}{n^T} \sum_{\mathbf{D}^T_-} \mathbb{1}_{[1-y_i^T(\langle \mathbf{w}, \mathbf{x}_i^T \rangle - b_2)>0]} \mathbf{x}_i^T y_i^T \\
&\quad + \frac{C_2}{n^C} \sum_{\mathbf{D}^C_+} \mathbb{1}_{[1+y_i^C(\langle \mathbf{w}, \mathbf{x}_i^C \rangle - b_1)>0]} \mathbf{x}_i^C y_i^C + \frac{C_1}{n^C} \sum_{\mathbf{D}^C_+} \mathbb{1}_{[1+y_i^C(\langle \mathbf{w}, \mathbf{x}_i^C \rangle - b_2)>0]} \mathbf{x}_i^C y_i^C \\
&\quad + \frac{C_1}{n^C} \sum_{\mathbf{D}^C_-} \mathbb{1}_{[1+y_i^C(\langle \mathbf{w}, \mathbf{x}_i^C \rangle - b_1)>0]} \mathbf{x}_i^C y_i^C + \frac{C_2}{n^C} \sum_{\mathbf{D}^C_-} \mathbb{1}_{[1+y_i^C(\langle \mathbf{w}, \mathbf{x}_i^C \rangle - b_2)>0]} \mathbf{x}_i^C y_i^C \\
&\quad + C_3 \frac{\partial S(\mathbf{D}^T, \mathbf{D}^C, \mathbf{w})}{\partial \mathbf{w}},
\end{aligned} \tag{8}$$

where

$$\begin{aligned}
&\frac{\partial S(\mathbf{D}^T, \mathbf{D}^C, \mathbf{w})}{\partial \mathbf{w}} \\
&\quad = \frac{2}{n^T n^C} \sum_{i=1}^{n^T} \sum_{j=1}^{n^C} \alpha |\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^C \rangle|^{\alpha-1} \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^C \rangle)(\mathbf{x}_i^T - \mathbf{x}_j^C)
\end{aligned}$$

$$- \frac{1}{(n^T)^2} \sum_{i=1}^{n^T} \sum_{j=1}^{n^T} \alpha |\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^T \rangle|^{\alpha-1} \text{sgn}(\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^T \rangle)(\mathbf{x}_i^T - \mathbf{x}_j^T)$$

$$- \frac{1}{(n^C)^2} \sum_{i=1}^{n^C} \sum_{j=1}^{n^C} \alpha |\langle \mathbf{w}, \mathbf{x}_i^C - \mathbf{x}_j^C \rangle|^{\alpha-1} \text{sgn}(\langle \mathbf{w}, \mathbf{x}_i^C - \mathbf{x}_j^C \rangle)(\mathbf{x}_i^C - \mathbf{x}_j^C). \qquad (9)$$

4.1 Averaged Stochastic Gradient Descent Algorithm for Székely Regularized USMVs

The Stochastic Gradient Descent algorithm typically works by picking random datapoints, computing the contribution of those points to the gradient and updating current weights with a decreasing update coefficient.

Notice, however, that each term in the Székely regularizer given in (5) operates on a pair of treatment datapoints and a pair of control datapoints. In order to apply a stochastic optimization algorithm to the problem we thus take, at each iteration, four randomly selected records, two from the treatment training set and two from control. The weight update is then computed based on four training points instead of one.

The algorithm is given in Fig. 2. The expressions $\partial l(\mathbf{w}, \mathbf{x}_i^T, y_i^T)/\partial \mathbf{w}$ and $\partial S(\mathbf{x}_i^T, \mathbf{x}_j^T, \mathbf{x}_k^C, \mathbf{x}_l^C, y_i^T, y_j^T, y_k^C, y_l^C, \mathbf{w})/\partial \mathbf{w}$ used in the algorithm will be given below. Notice that in step 10 we take the average of the weight vectors $\mathbf{w}_t$ obtained during all steps of the algorithm. This is the so called Polyak-Ruppert averaging [1, 7, 9] which improves the convergence properties of the algorithm.

In order to provide the expressions for $\partial l/\partial \mathbf{w}$ and $\partial S/\partial \mathbf{w}$ used in the algorithm, as well as to prove its convergence, we first need to compute the subgradient of the target risk function (9) for each random sample. Since we are dealing with pairs of treatment

```
1.  $\mathbf{w}_0 = 0$
2.  For $t \leftarrow 1, 2, \ldots$
3.      Draw two samples $(\mathbf{x}_i^T, y_i^T)$, $(\mathbf{x}_j^T, y_j^T)$ uniformly at random from $\mathbf{D}^T$
4.      Draw two samples $(\mathbf{x}_k^C, y_k^C)$, $(\mathbf{x}_l^C, y_l^C)$ uniformly at random from $\mathbf{D}^C$
5.      $\mathbf{g} \leftarrow \mathbf{w}_{t-1} + \frac{1}{2}\frac{\partial l(\mathbf{w}_{t-1}, \mathbf{x}_i^T, y_i^T)}{\partial \mathbf{w}_{t-1}} + \frac{1}{2}\frac{\partial l(\mathbf{w}_{t-1}, \mathbf{x}_j^T, y_j^T)}{\partial \mathbf{w}_{t-1}}$
6.      $\mathbf{g} \leftarrow \mathbf{g} + \frac{1}{2}\frac{\partial l(\mathbf{w}_{t-1}, \mathbf{x}_k^C, y_k^C)}{\partial \mathbf{w}_{t-1}} + \frac{1}{2}\frac{\partial l(\mathbf{w}_{t-1}, \mathbf{x}_l^C, y_l^C)}{\partial \mathbf{w}_{t-1}}$
7.      $\mathbf{g} \leftarrow \mathbf{g} + C_3 \frac{\partial S(\mathbf{x}_i^T, \mathbf{x}_j^T, \mathbf{x}_k^C, \mathbf{x}_l^C, y_i^T, y_j^T, y_k^C, y_l^C, \mathbf{w}_{t-1})}{\partial \mathbf{w}_{t-1}}$
8.      $\gamma_t \leftarrow \frac{1}{\sqrt{t}}$
9.      $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \gamma_t \mathbf{g}$
10.     $\mathbf{w}^a \leftarrow \frac{1}{t}\sum_{t'=1}^{t} \mathbf{w}_{t'}$
11.     If converged:
12.         Return $\mathbf{w}^a$
```

Fig. 2 The Averaged Stochastic Gradient descent algorithm for Székely regularized uplift support vector machines

and control points, each sample will involve four data records: $\mathbf{x}_i^T, \mathbf{x}_j^T, \mathbf{x}_k^C, \mathbf{x}_l^C$ and their corresponding class values $y_i^T, y_j^T, y_k^C, y_l^C$. The subgradient of the risk for the given sample is given by the following equation

$$\begin{aligned}
&\frac{\partial R(\mathbf{x}_i^T, \mathbf{x}_j^T, \mathbf{x}_k^C, \mathbf{x}_l^C, y_i^T, y_j^T, y_k^C, y_l^C, \mathbf{w})}{\partial \mathbf{w}} \\
&\quad = \mathbf{w} + \frac{1}{2}\frac{\partial l(\mathbf{w}, \mathbf{x}_i^T, y_i^T)}{\partial \mathbf{w}} + \frac{1}{2}\frac{\partial l(\mathbf{w}, \mathbf{x}_j^T, y_j^T)}{\partial \mathbf{w}} \\
&\quad + \frac{1}{2}\frac{\partial l(\mathbf{w}, \mathbf{x}_k^C, y_k^C)}{\partial \mathbf{w}} + \frac{1}{2}\frac{\partial l(\mathbf{w}, \mathbf{x}_l^C, y_l^C)}{\partial \mathbf{w}} \\
&\quad + C_3 \frac{\partial S(\mathbf{x}_i^T, \mathbf{x}_j^T, \mathbf{x}_k^C, \mathbf{x}_l^C, y_i^T, y_j^T, y_k^C, y_l^C, \mathbf{w})}{\partial \mathbf{w}},
\end{aligned} \tag{10}$$

where the parts resulting from differentiating the hinge loss are

$$\frac{\partial l(\mathbf{w}, \mathbf{x}, y)}{\partial \mathbf{w}} = \mathbf{x}y \cdot \begin{cases} -C_1 \mathbb{1}_{[1-y(\langle \mathbf{w}, \mathbf{x}\rangle - b_1)>0]} - C_2 \mathbb{1}_{[1-y(\langle \mathbf{w}, \mathbf{x}\rangle - b_2)>0]} & \text{if } (\mathbf{x}, y) \in \mathbf{D}_+^T \\ -C_2 \mathbb{1}_{[1-y(\langle \mathbf{w}, \mathbf{x}\rangle - b_1)>0]} - C_1 \mathbb{1}_{[1-y(\langle \mathbf{w}, \mathbf{x}\rangle - b_2)>0]} & \text{if } (\mathbf{x}, y) \in \mathbf{D}_-^T \\ C_1 \mathbb{1}_{[1+y(\langle \mathbf{w}, \mathbf{x}\rangle - b_1)>0]} + C_2 \mathbb{1}_{[1+y(\langle \mathbf{w}, \mathbf{x}\rangle - b_2)>0]} & \text{if } (\mathbf{x}, y) \in \mathbf{D}_-^C \\ C_2 \mathbb{1}_{[1+y(\langle \mathbf{w}, \mathbf{x}\rangle - b_1)>0]} + C_1 \mathbb{1}_{[1+y(\langle \mathbf{w}, \mathbf{x}\rangle - b_2)>0]} & \text{if } (\mathbf{x}, y) \in \mathbf{D}_+^C \end{cases} \tag{11}$$

and the part for the subgradient of the Székely regularizer is

$$\begin{aligned}
&\frac{\partial S(\mathbf{x}_i^T, \mathbf{x}_j^T, \mathbf{x}_k^C, \mathbf{x}_l^C, y_i^T, y_j^T, y_k^C, y_l^C, \mathbf{w})}{\partial \mathbf{w}} \\
&\quad = \frac{\alpha}{2}\big[|\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_k^C\rangle|^{\alpha-1}\operatorname{sgn}(\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_k^C\rangle)(\mathbf{x}_i^T - \mathbf{x}_k^C) \\
&\qquad + |\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_l^C\rangle|^{\alpha-1}\operatorname{sgn}(\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_l^C\rangle)(\mathbf{x}_i^T - \mathbf{x}_l^C) \\
&\qquad + |\langle \mathbf{w}, \mathbf{x}_j^T - \mathbf{x}_k^C\rangle|^{\alpha-1}\operatorname{sgn}(\langle \mathbf{w}, \mathbf{x}_j^T - \mathbf{x}_k^C\rangle)(\mathbf{x}_j^T - \mathbf{x}_k^C) \\
&\qquad + |\langle \mathbf{w}, \mathbf{x}_j^T - \mathbf{x}_l^C\rangle|^{\alpha-1}\operatorname{sgn}(\langle \mathbf{w}, \mathbf{x}_j^T - \mathbf{x}_l^C\rangle)(\mathbf{x}_j^T - \mathbf{x}_l^C)\big] \\
&\quad - \alpha|\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^T\rangle|^{\alpha-1}\operatorname{sgn}(\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^T\rangle)(\mathbf{x}_i^T - \mathbf{x}_j^T) \\
&\quad - \alpha|\langle \mathbf{w}, \mathbf{x}_k^C - \mathbf{x}_l^C\rangle|^{\alpha-1}\operatorname{sgn}(\langle \mathbf{w}, \mathbf{x}_k^C - \mathbf{x}_l^C\rangle)(\mathbf{x}_k^C - \mathbf{x}_l^C).
\end{aligned}$$

A necessary condition for the Stochastic Gradient Descent algorithm to converge is that the expectation (taken over the randomly sampled vectors) of the subgradient (10) be equal to the subgradient computed on the full dataset given in (8). Since in the algorithm given in Fig. 2 we are using four randomly sampled data points we need to take the expectation over all of them.

We will denote the expectation over $(\mathbf{x}_i^T, y_i^T)$ by $\mathbb{E}_i^T[\cdot]$, analogous notation will be used for expectations over records in the control group. Notice that, since the records in the stochastic optimization algorithm are chosen uniformly at random, we have

$$\mathbb{E}_i^T[f(\mathbf{x}_i^T, y_i^T)] = \frac{1}{n^T}\sum_{i=1}^{n^T} f(\mathbf{x}_i^T, y_i^T), \quad \mathbb{E}_k^C[f(\mathbf{x}_i^C, y_i^C)] = \frac{1}{n^C}\sum_{i=1}^{n^C} f(\mathbf{x}_i^C, y_i^C). \tag{12}$$

Further, denote by $\mathbb{E}[\cdot]$ the expectation over all four randomly chosen samples $(\mathbf{x}_i^T, y_i^T)$, $(\mathbf{x}_j^T, y_j^T)$, $(\mathbf{x}_k^C, y_k^C)$, $(\mathbf{x}_l^C, y_l^C)$, i.e.

$$\mathbb{E}[\cdot] = \mathbb{E}_i^T\mathbb{E}_j^T\mathbb{E}_k^C\mathbb{E}_l^C[\cdot].$$

Let us now compute, term by term, the expectation of the subgradient given in (10). Clearly $\mathbb{E}\mathbf{w} = \mathbf{w}$. Also, using (11) and (12) we get

$$\begin{aligned}
\mathbb{E}\frac{\partial l(\mathbf{w}, \mathbf{x}_i^T, y_i^T)}{\partial \mathbf{w}} &= \mathbb{E}_i^T\frac{\partial l(\mathbf{w}, \mathbf{x}_i^T, y_i^T)}{\partial \mathbf{w}}\\
&= -\frac{1}{n^T}\sum_{i=1}^{n^T}\mathbf{x}_i^T y_i^T \begin{cases} C_1\mathbb{1}_{[1-y_i^T(\langle\mathbf{w},\mathbf{x}_i^T\rangle-b_1)>0]} \\ \quad +C_2\mathbb{1}_{[1-y_i^T(\langle\mathbf{w},\mathbf{x}_i^T\rangle-b_2)>0]} & \text{if } (\mathbf{x}_i^T, y_i^T)\in\mathbf{D}_+^T\\ C_2\mathbb{1}_{[1-y_i^T(\langle\mathbf{w},\mathbf{x}_i^T\rangle-b_1)>0]} \\ \quad +C_1\mathbb{1}_{[1-y_i^T(\langle\mathbf{w},\mathbf{x}_i^T\rangle-b_2)>0]} & \text{if } (\mathbf{x}_i^T, y_i^T)\in\mathbf{D}_-^T\end{cases}\\
&= -\frac{C_1}{n^T}\sum_{\mathbf{D}_+^T}\mathbb{1}_{[1-y_i^T(\langle\mathbf{w},\mathbf{x}_i\rangle-b_1)>0]}\mathbf{x}_i^T y_i^T - \frac{C_2}{n^T}\sum_{\mathbf{D}_+^T}\mathbb{1}_{[1-y_i^T(\langle\mathbf{w},\mathbf{x}_i\rangle-b_2)>0]}\mathbf{x}_i^T y_i^T\\
&\quad - \frac{C_2}{n^T}\sum_{\mathbf{D}_-^T}\mathbb{1}_{[1-y_i^T(\langle\mathbf{w},\mathbf{x}_i^T\rangle-b_1)>0]}\mathbf{x}_i^T y_i^T - \frac{C_1}{n^T}\sum_{\mathbf{D}_-^T}\mathbb{1}_{[1-y_i^T(\langle\mathbf{w},\mathbf{x}_i^T\rangle-b_2)>0]}\mathbf{x}_i^T y_i^T.
\end{aligned}$$

We now move to computing the expectation of the subgradient of the Székely regularizer. Note that

$$\begin{aligned}
&\mathbb{E}|\langle\mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_k^C\rangle|^{\alpha-1}\mathrm{sgn}(\langle\mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_k^C\rangle)(\mathbf{x}_i^T - \mathbf{x}_k^C)\\
&\qquad = \mathbb{E}_i^T\mathbb{E}_k^C|\langle\mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_k^C\rangle|^{\alpha-1}\mathrm{sgn}(\langle\mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_k^C\rangle)(\mathbf{x}_i^T - \mathbf{x}_k^C)\\
&\qquad = \frac{1}{n^T n^C}\sum_{i=1}^{n^T}\sum_{k=1}^{n^C}|\langle\mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_k^C\rangle|^{\alpha-1}\mathrm{sgn}(\langle\mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_k^C\rangle)(\mathbf{x}_i^T - \mathbf{x}_k^C).
\end{aligned}$$

By symmetry, the three other pairs of treatment and control points lead to the same expected value. Similarly

$$\mathbb{E}|\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^T \rangle|^{\alpha-1}\mathrm{sgn}(\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^T \rangle)(\mathbf{x}_i^T - \mathbf{x}_j^T)$$
$$= \frac{1}{(n^T)^2}\sum_{i=1}^{n^T}\sum_{j=1}^{n^T} |\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^T \rangle|^{\alpha-1}\mathrm{sgn}(\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^T \rangle)(\mathbf{x}_i^T - \mathbf{x}_j^T).$$

The expression for the pair of control points is analogous. Finally we get

$$\mathbb{E}\frac{\partial S(\mathbf{x}_i^T, \mathbf{x}_j^T, \mathbf{x}_k^C, \mathbf{x}_l^C, y_i^T, y_j^T, y_k^C, y_l^C, \mathbf{w})}{\partial \mathbf{w}}$$
$$= 2\alpha\frac{1}{n^T n^C}\sum_{i=1}^{n^T}\sum_{k=1}^{n^C} |\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_k^C \rangle|^{\alpha-1}\mathrm{sgn}(\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_k^C \rangle)(\mathbf{x}_i^T - \mathbf{x}_k^C)$$
$$- \alpha\frac{1}{(n^T)^2}\sum_{i=1}^{n^T}\sum_{j=1}^{n^T} |\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^T \rangle|^{\alpha-1}\mathrm{sgn}(\langle \mathbf{w}, \mathbf{x}_i^T - \mathbf{x}_j^T \rangle)(\mathbf{x}_i^T - \mathbf{x}_j^T)$$
$$- \alpha\frac{1}{(n^C)^2}\sum_{i=1}^{n^C}\sum_{j=1}^{n^C} |\langle \mathbf{w}, \mathbf{x}_i^C - \mathbf{x}_j^C \rangle|^{\alpha-1}\mathrm{sgn}(\langle \mathbf{w}, \mathbf{x}_i^C - \mathbf{x}_j^C \rangle)(\mathbf{x}_i^C - \mathbf{x}_j^C).$$

Combining the above results we get exactly the subgradient of the risk which is minimized by Székely regularized Uplift Support Vector Machines given in (8) and (9).

Therefore the necessary condition for convergence is satisfied. For sufficiency, let us examine the properties of the optimization problem (6). Notice first, that although the term $\frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle$ is strongly convex and the remaining terms are convex, the Székely penalty term is not. Therefore the optimized function need not be convex and we cannot guarantee global convergence. Suppose that there exists a bound D such that $\|\mathbf{w}^*\| \leq D$ and $\|\mathbf{w}_t\| \leq D$ for all iteration steps t. Note that the risk function is Lipschitz continuous on any closed region of the parameter space. It follows that the subgradient of R is bounded throughout the algorithm and the convergence is guaranteed based on results given in [7, Sect. 11.0] for $\gamma_t = Ct^{-\frac{1}{2}}$. The constant C was chosen to be 1 in our implementation.

Let us now briefly comment on the existence of the bound D. Without additional assumptions we cannot formally guarantee that at every iteration we have $\|\mathbf{w}_t\| \leq D$. To obtain such guarantees, an extra step can be added to Algorithm 2, which, after each iteration, projects $\mathbf{w}_t$ onto a ball of some radius D [7]. In practice we saw no convergence problems and the extra step was not necessary.

If we make an additional assumption that the Székely penalty S is locally convex around the minimum we can guarantee fast convergence rates. Since $\frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle$ is strongly convex and a sum of a convex and a strongly convex function is strongly convex, the risk $R(\mathbf{w})$ given in (6) becomes strongly convex. The convergence rate is then $O(t^{-1})$ for $\gamma_t = Ct^{-\frac{1}{2}}$ following the results in [1, Theorem 3]. The constant C was chosen to be 1 in our implementation.

Note that the step size used guarantees convergence for non strongly convex functions and fast convergence for strongly convex ones.

5 Experimental Evaluation

In this section we will present an experimental evaluation of the proposed approach. We begin with a general discussion on evaluating uplift models, then describe our approach to testing under treatment assignment bias, and finally present the results of experiments.

5.1 Evaluating Uplift Models

Evaluating uplift models is more difficult than evaluating traditional classifiers. For each case we only know the outcome after the action was applied or when the action was not taken, never both. Therefore, we cannot decide whether specific cases have been correctly classified or not. This is known as the fundamental problem of causal inference [3].

Let us now discuss so called uplift curves used to graphically analyze the performance of uplift models. One type of curves used to assess standard classifiers are lift curves (also known as cumulative gains curves or cumulative accuracy profiles), where the x axis corresponds to the number of cases targeted and the y axis to the number of successes captured by the model. An *uplift curve* is computed by subtracting the lift curve obtained on the control test set from the lift curve obtained on the treatment test set. Both curves are generated using the same uplift model. The number of successes on the y axes is expressed as a percentage of the total population which guarantees that the curves can be meaningfully subtracted.

The interpretation of the uplift curve is as follows: on the x axis we choose the percentage of the population on which the action is to be performed and on the y axis we read the difference between the success rates in the treatment and control groups. The value at $x = 100\,\%$ gives the gain in success probability from treating the whole population. A diagonal line corresponds to random selection. The Area Under the Uplift Curve (AUUC) can be used as a simple summary. In this paper we subtract the area under the diagonal line from this value in order to obtain more meaningful numbers. More details on evaluating uplift models and on uplift curves can be found in [11, 15].

Since there are now two test sets (treatment and control) procedures such as crossvalidation are performed independently on them. In this paper we use ten times ten-fold crossvalidation to obtain the uplift curves.

5.2 The Right Heart Catheterization Dataset

The right heart catheterization dataset [5] contains data about 5735 patients admitted to hospitals in serious condition. 2184 of them were subjected to the right heart catheterization procedure (RHC) and constitute the treatment group; the remaining 3551 did not receive the procedure and are the controls. The data does not come from a randomized study, the application of RHC was decided based on patients' condition, so the group selection is biased, in fact, it was done retrospectively based on historical data. Because of this characteristics, as well as its relatively large size, the dataset is ideal to test our algorithm.

The class variable was the attribute `Death` denoting patient death during the first 180 days after hospital admission. Patient survival was considered the positive outcome. To avoid information leaks we removed other outcome related variables such as date of death or date of last contact.

The predictive attributes describe various characteristics of the patient such as age, sex, education, income, medical insurance, the disease the patient suffers from. Also present are results of diagnostics performed at admission such as blood pressure, temperature, results of blood tests, and various scores describing the severity of patient's condition.

5.3 Testing Methodology—Correcting Group Selection Bias

Testing the performance of the models was, however, more challenging than in case of randomized controlled trials. As discussed in Sect. 5.1 testing uplift models is based on an assumption that groups of treatment and control records with similar scores are indeed similar. Unfortunately, this is usually not the case for biased treatment selection.

In order to test the model's predictions we thus had to correct the bias in the test sets. In practice, such corrections are typically achieved using so called propensity scores [14]. A propensity score is the probability that a given patient, described by a feature vector $\mathbf{x}$, will be assigned to the treatment group. There are several ways propensity scores can be used to correct for nonrandom group assignment. In this paper we are going to use inverse probability of treatment weighting (IPTW) [13].

The IPTW method assigns to each treatment group record a weight inversely proportional to the probability that a record with those characteristics is selected for treatment. This way, cases to which treatment is applied disproportionately often are given lower weights and underrepresented cases higher weights. The control group records are, analogously, given weights proportional to the inverse of the probability that a record with a given feature vector is not given the treatment. Note that for a randomized controlled trial all records within a group receive equal weights.

To summarize, our testing procedure works as follows: we build a model with nonrandom treatment assignment using the Székely penalty term to correct for the

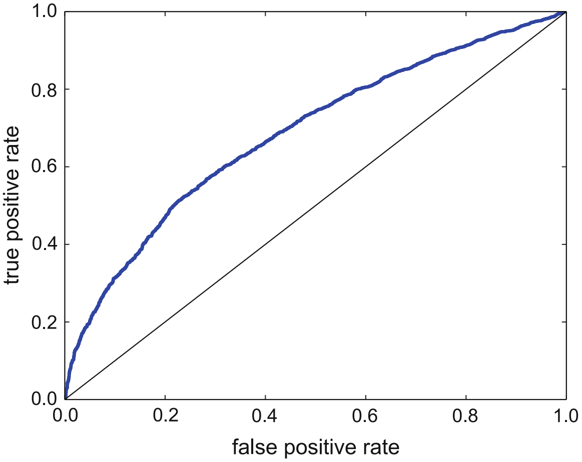


Fig. 3 ROC curve for the propensity score model

bias, then we test the model on treatment and control test sets on which the bias has been corrected using the inverse probability of treatment weighting. Since different bias correction procedures are used for model construction and testing, we believe that it is less likely that the estimated model performance is a result of an uncorrected treatment assignment bias.

In order to use the IPTW procedure one needs to know the probability of treatment assignment conditional on patients' characteristics. Unfortunately, this probability is usually unknown and needs to be estimated. Here, we use a logistic regression model trained on full data before crossvalidation splits. The ROC curve for the model is shown in Fig. 3 (area under the ROC curve is 0.686). It can be seen that the model is able to predict reasonably well whether a given patient will receive the RHC procedure. One can conclude that treatment assignment is indeed seriously biased.

5.4 Experimental Results

We will now present the experimental results. Figure 4 shows uplift curves drawn for several values of the Székely penalty coefficient C_3. All experiments were performed for $C_1 = C_2 = 1$, only the value of C_3 was changed. Ten times ten-fold crossvalidation was used to obtain the curves. The curves are drawn based on data weighted using inverse probability of treatment weighting (IPTW) to correct treatment assignment bias. More detailed data on areas under the uplift curves are given in the second row of Table 1.

Overall the treatment is not effective and the application of the right heart catheterization procedure seems to decrease patients chances of survival. This is in line with the findings presented in [5].

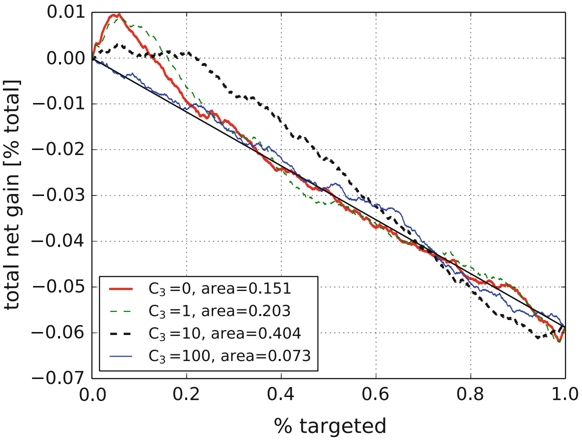


Fig. 4 Uplift curves for Uplift SVM models with different Székely penalty coefficients

Table 1 The influence of the Székely penalty coefficient C_3 on the area under the uplift curve and the differences between scores in the treatment and control groups

C_3 penalty coefficient	0	0.01	0.1	1	10	100
AUUC	0.1505	0.1511	0.1559	0.2030	0.4035	0.0727
Difference between score means	0.1051	0.1050	0.1036	0.0839	0.0202	0.004
Kolmogorov-Smirnov statistic	0.3436	0.3433	0.3411	0.3026	0.1103	0.0320

It can be seen that without the Székely correction ($C_3 = 0$) the curve follows the diagonal line corresponding to a model assigning treatment at random, except for the 20 % of highest scored cases for which RHC does indeed bring improvement in survival rate over random selection.

With increasing values of the Székely penalty coefficient C_3, the area under the uplift curve is steadily increasing, up to $C_3 = 100$ where the performance rapidly drops. This shows that the application of the Székely penalty does indeed improve model performance under treatment assignment bias.

The best performance is achieved for $C_3 = 10$, and Fig. 4 shows that this particular model achieves good performance over a wide range of scores, bringing improvement over random selection for about 75 % of the population. The area under the uplift curve is more than two and a half times better than for the unregularized model.

The drop in performance for very high value of the regularization parameter is typical for regularized models in general: too high a penalty leads to the model ignoring the data and focusing only on the regularization term.

To further analyze the effect of the Székely penalty on model behavior we analyze the distributions of model scores in treatment and control groups. Figure 5 shows

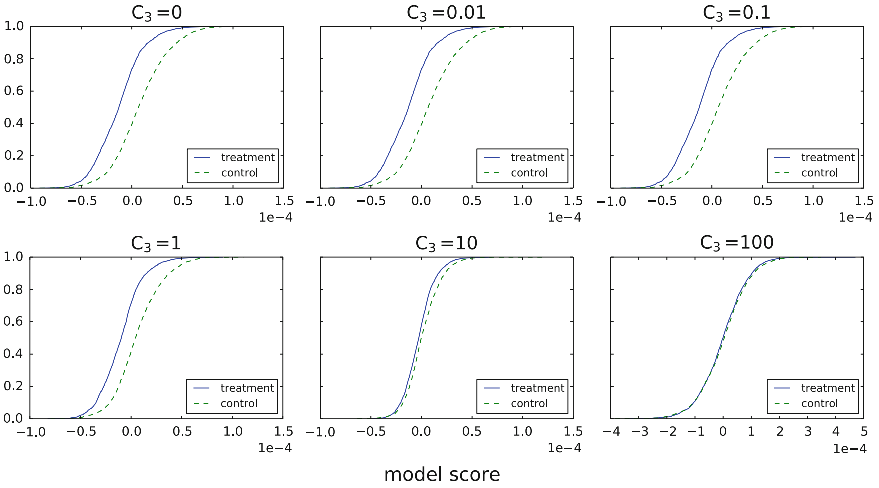


Fig. 5 Empirical cumulative distribution functions of scores in the treatment and control groups for various Székely regularization coefficient values

the empirical cumulative distribution functions of model score distributions in the treatment and control groups for various strength of the Székely regularization term. The charts were obtained on a single repetition of the ten fold crossvalidation. Additionally, we computed two types of statistics summarizing the discrepancies: the difference between score means in the two groups and the Kolmogorov-Smirnov statistic, i.e. the maximum difference between the empirical cumulative distribution functions of the two groups. The summary statistics are given in the third and fourth rows of Table 1 and are shown graphically in Fig. 6.

It can be seen that for the unregularized model, the distributions of scores in both groups differ significantly. The score means differ by about 0.1, which is a fairly large value since the scores range roughly from -0.5 to 0.5. The value of the Kolmogorov-Smirnov statistic is almost 0.35.

When the Székely penalty increases, the distributions become closer to each other. For $C_3 = 0.01$ and $C_3 = 0.1$ the decrease is tiny but noticeable and is accompanied by a tiny but noticeable improvement in the area under the uplift curve. When $C_3 = 1$ the score distributions already come much closer to each other with the difference between means decreasing to about 0.084; at the same time AUUC increased by about 35 % with respect to the unregularized model. A further tenfold increase of the penalty coefficient makes the distributions very similar; the difference in means is just 0.02 and the Kolmogorov-Smirnov statistic just 0.11. The AUUC is 2.68 times higher than for the unregularized model.

A further tenfold increase in the value of C_3 makes the score distributions in the treatment and control groups practically identical, however the regularization is too strong and the model no longer correctly predicts for whom the RHC procedure is beneficial. In fact its predictions are not better than random assignments.

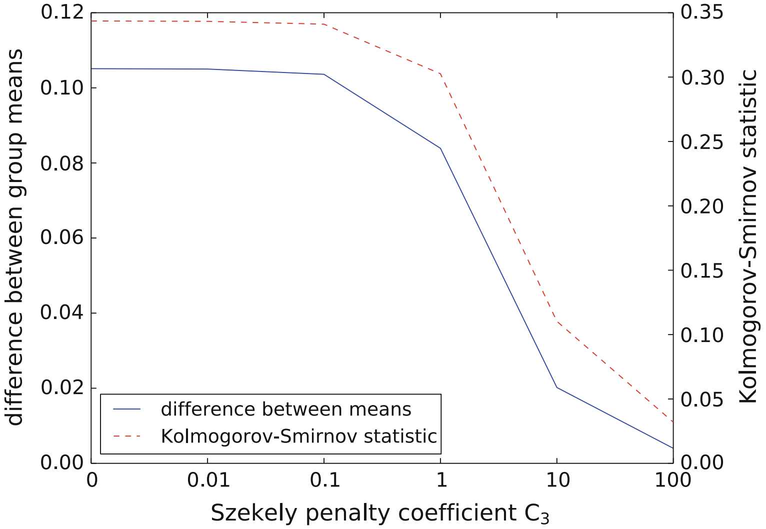


Fig. 6 Statistics summarizing the differences between score distributions in the treatment and control groups for various Székely regularization coefficient values

Let us now summarize our experimental findings. First, it was shown that the unregularized model behaves poorly after treatment assignment bias correction is applied. Moreover, it produces significantly different scores in the treatment and control groups likely modeling not the real causal impact of the action but the differences in group assignment. As the Székely penalty term increased, the differences between scores the model assigns to treatment and control records became much smaller, accompanied by large improvements in model performance. One can thus conclude that using the Székely penalty term does indeed reduce model's susceptibility to treatment assignment bias, proving the main claim of the paper.

6 Conclusions

We have presented a regularization method which corrects the behavior of uplift models under nonrandomized treatment assignment. The approach is based on an energy distance proposed by Székely and Rizzo which offers a practical way of ensuring similarity of model scores in the treatment and control datasets. Experiments performed on the right heart catheterization dataset confirm the usefulness of the proposed approach.

Acknowledgments This work was supported by Research Grant no. N N516 414938 of the Polish Ministry of Science and Higher Education (Ministerstwo Nauki i Szkolnictwa Wyższego) from research funds for the period 2010–2014. Ł.Z. was co-funded by the European Union from resources of the European Social Fund. Project POKL 'Information technologies: Research and their interdisciplinary applications', Agreement UDA-POKL.04.01.01-00-051/10-00.

References

1. Bach F, Moulines E (2011) Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: Proceedings of advances in neural information processing systems 24 (NIPS 2011)
2. Guelman L, Guillén M, Pérez-Marín AM (2012) Random forests for uplift modeling: an insurance customer retention case. In: Modeling and simulation in engineering, economics and management. Lecture notes in business information processing (LNBIP), vol 115. Springer, pp. 123–133
3. Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81(396):945–960
4. Jaśkowski M, Jaroszewicz S (2012) Uplift modeling for clinical trial data. In: ICML 2012 workshop on machine learning for clinical data analysis, Edinburgh, June 2012
5. Jr Connors AF, Speroff T, Dawson NV et al (1996) The effectiveness of right heart catheterization in the initial care of critically ill patients. JAMA 276(11):889–897
6. Koronacki J, Ćwik J (2008) Statystyczne systemy uczące się. Exit, Warsaw (In Polish)
7. Kushner HJ, Yin GG (2003) Stochastic approximation and recursive algorithms and applications. Springer
8. Kuusisto F, Costa VS, Nassif H, Burnside E, Page D, Shavlik J (2014) Support vector machines for differential prediction. In: ECML-PKDD
9. Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. SIAM J Control Optim 30(4):838–855
10. Radcliffe NJ, Surry PD (1999) Differential response analysis: Modeling true response by isolating the effect of a single action. In: Proceedings of credit scoring and credit control VI. Credit Research Centre, University of Edinburgh Management School
11. Radcliffe NJ, Surry PD (2011) Real-world uplift modelling with significance-based uplift trees. Portrait Technical Report TR-2011-1, Stochastic Solutions
12. Robins J, Rotnitzky A (2004) Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. Biometrika 91(4):763–783
13. Rosenbaum PR (1987) Model-based direct adjustment. J Am Stat Assoc 82(398):387–394
14. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55
15. Rzepakowski P, Jaroszewicz S (2010) Decision trees for uplift modeling. In: Proceedings of the 10th IEEE international conference on data mining (ICDM), Sydney, Australia, pp. 441–450 Dec 2010
16. Rzepakowski P, Jaroszewicz S (2012) Decision trees for uplift modeling with single and multiple treatments. Knowl Inf Syst 32:303–327 August
17. Sołtys M, Jaroszewicz S, Rzepakowski P (2014) Ensemble methods for uplift modeling. Data mining and knowledge discovery, pp. 1–29 (online first)
18. Szekely GJ, Rizzo ML (2004) Testing for equal distributions in high dimension. Interstat, Nov 2004
19. Szekely GJ, Rizzo ML (2005) Hierarchical clustering via joint between-within distances: extending ward's minimum variance method. J Classif 22(2):151–183
20. Szekely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. Ann Stat 35(6):2769–2794
21. Vansteelandt S, Goetghebeur E (2003) Causal inference with generalized structural mean models. J R Stat Soc B 65(4):817–835
22. Zaniewicz L, Jaroszewicz S (2013) Support vector machines for uplift modeling. In: The first IEEE ICDM workshop on causal discovery (CD 2013), Dallas, Dec 2013

Dominance-Based Rough Set Approach to Multiple Criteria Ranking with Sorting-Specific Preference Information

Miłosz Kadziński, Roman Słowiński and Marcin Szeląg

Abstract A novel multiple criteria decision aiding method is proposed, that delivers a recommendation characteristic for ranking problems but employs preference information typical for sorting problems. The method belongs to the category of ordinal regression methods: it starts with preference information provided by the Decision Maker (DM) in terms of decision examples, and then builds a preference model that reproduces these exemplary decisions. The ordinal regression is analogous to inductive learning of a model that is true in the closed world of data where it comes from. The sorting examples show an assignment of some alternatives to pre-defined and ordered quality classes. Although this preference information is purely ordinal, the number of quality classes separating two assigned alternatives is meaningful for an ordinal intensity of preference. Using an adaptation of the Dominance-based Rough Set Approach (DRSA), the method builds from this information a decision rule preference model. This model is then applied on a considered set of alternatives to finally rank them from the best to the worst. The decision rule preference model resulting from DRSA is able to represent the preference information about the ordinal intensity of preference without converting this information into a cardinal scale. Moreover, the decision rules can be interpreted straightforwardly by the DM, facilitating her understanding of the feedback between the preference information and the preference model. An illustrative case study performed in this paper supports this claim.

Keywords Decision analysis · Preference learning · Ranking · Dominance-based rough set approach · Decision rules · Assignment examples

M. Kadziński · R. Słowiński (✉) · M. Szeląg
Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
e-mail: roman.slowinski@cs.put.poznan.pl

M. Kadziński
e-mail: milosz.kadzinski@cs.put.poznan.pl

M. Szeląg
e-mail: marcin.szelag@cs.put.poznan.pl

R. Słowiński
Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_9

1 Introduction

Decision problems considered in Multiple Criteria Decision Aiding (MCDA) concern a set of alternatives evaluated on a consistent family of criteria. MCDA gives the decision makers some tools and methods for structuring the problem, preference handling, and carrying forward the process of decision making. Taking into account the type of expected results, the term decision can be interpreted in different ways. Generally, in MCDA we distinguish three types of decision problems: choice, ranking, and sorting.

In choice problems, one aims at selecting a small subset of potentially best alternatives. In ranking problems, alternatives should be ordered from the best to the worst. Finally, the sorting problem is about assigning the alternatives to some predefined and ordered classes. When considering multiple conflicting criteria, arriving at a recommendation for each type of decision problems requires the use of some particular decision aiding method [19].

Each MCDA method is distinguished by the type of admitted preference information, ways of constructing and exploiting the preference model, and techniques used to work out a recommendation. Usually, these methods are designed for dealing with either ranking and choice or sorting (ordinal classification) problems.

In this paper, we introduce a novel MCDA method that delivers recommendation characteristic for ranking problems but makes use of preference information that is typical for sorting problems. This method employs an adaptation of the Dominance-based Rough Set Approach (DRSA) (see [8, 11, 12, 21, 22]). Given the preference information in terms of class assignments (sorting) of some reference alternatives, it builds a decision rule preference model. This model is then applied on a set of alternatives to be ranked, yielding a recommendation in terms of a weak order of these alternatives.

The method proposed belongs to the category of ordinal regression methods that start with preference information provided by the Decision Maker (DM) in terms of decision examples, and then build a mathematical model that replicates these exemplary decisions. For this ability, the model is called DM's preference model.

The motivation behind the proposed approach to multiple criteria ranking is twofold:

- the preference information provided by the DM in terms of sorting examples permits to express the intensity of preference between alternatives in a purely ordinal way, such that intensity of preference of a over b is comparable to that of c over d only if the interval of classes between the assignment of a and b includes or is included in the interval of c and d; otherwise the intensities are non-comparable;
- the decision rule preference model resulting from DRSA is able to express the preference relation with the above meaning of the ordinal intensity without any transformation of the input preference information; moreover, the decision rules can be interpreted straightforwardly by the DM, facilitating her understanding of the feedback between the preference information and the preference model.

Let us comment on these motivations in more detail. First, let us observe that when the final aim is to construct a ranking of all considered alternatives, the preference information has often the form of pairwise comparisons of some or all considered alternatives. This is quite natural, because position of each alternative in the ranking depends on result of its comparison with other alternatives. These pairwise comparisons often admit a multi-graded preference relation, expressing an intensity of preference.

For example, in the Analytical Hierarchy Process [20], the DM is supposed to compare pairwise all considered alternatives and express the intensity of preference on a pre-defined cardinal ratio scale. In the MACBETH method [1], all pairs of alternatives are assigned to some ordered classes of preference intensity and, finally, a cardinal intensity scale concordant with these assignments is computed. Some other methods do not require more from the DM than an ordinal expression of preference intensity, like "*a* is preferred to *b* at least as strongly as *c* is preferred to *d*", and obtain in consequence a single-graded quaternary relation in the set of pairs of alternatives; this is the case of the GRIP method [7], which builds a set of general additive value functions that replicate the ordinal preference information, and provides at the output necessary and possible quaternary relations that contribute to construction of necessary and possible rankings, respectively. All above methods use a value function preference model. Observe that rankings established by a value function permit to speak about intensity of preference between alternatives in the ranking, as the scale of the value function is an interval scale.

MCDA methods that use outranking relation preference model do not consider intensity of preference either in the input preference information, or in the resulting ranking, which has an ordinal character.

Methods based on logical representation of preferences in terms of monotonic decision rules, like DRSA, are able to process preference information with specified intensity of preference [8]. In this case, the pairwise comparisons of some reference alternatives get a degree of intensity of preference assigned by the DM. These degrees are linearly ordered, so that DRSA can approximate upward and downward unions of the degrees. Decision rules induced from these approximations suggest assignment of a pair of alternatives to a preference relation having at least or at most a specified degree of intensity. Application of these rules on a set of alternatives leads to a fuzzy preference graph, whose exploitation with a weighted fuzzy net flow score procedure leads to a final ranking. A difference of positions in this ranking does not have the meaning of intensity of preference, which is similar to rankings obtained by methods using as preference model an outranking relation.

Experience indicates, however, that answering the questions about the intensity of preference between two alternatives in cardinal terms requires too big cognitive effort on the part of the DM. Facilitating the DM's elicitation of the intensity of preference for pairwise comparisons is thus the first motivation of this paper. The method proposed in this paper permits the DM to express the intensity of preference between reference alternatives in a purely ordinal way, as assignments of alternatives to pre-defined and ordered quality classes. The order of these classes has no cardinal meaning, so that the number of classes separating two assigned alternatives is not meaningful for intensity of preference. The only comparison of intensities of preference is possible when an interval of classes for alternatives a, b includes or is included in the interval of classes for alternatives c, d; precisely, the following eight situations of comparability for intensities of preference may occur:

- $(a, b) \succeq (c, d)$ if $a \succeq b$, $c \succeq d$, and the interval of classes for a, b includes the interval of classes for c, d;
- $(a, b) \succeq (c, d)$ if $a \succeq b$, $c \preceq d$, and the interval of classes for a, b includes the interval of classes for c, d;
- $(a, b) \succeq (c, d)$ if $a \succeq b$, $c \preceq d$, and the interval of classes for a, b is included in the interval of classes for c, d;
- $(a, b) \succeq (c, d)$ if $a \preceq b$, $c \preceq d$, and the interval of classes for a, b is included in the interval of classes for c, d;
- $(a, b) \preceq (c, d)$ if $a \preceq b$, $c \succeq d$, and the interval of classes for a, b includes the interval of classes for c, d;
- $(a, b) \preceq (c, d)$ if $a \preceq b$, $c \preceq d$, and the interval of classes for a, b includes the interval of classes for c, d;
- $(a, b) \preceq (c, d)$ if $a \succeq b$, $c \succeq d$, and the interval of classes for a, b is included in the interval of classes for c, d;
- $(a, b) \preceq (c, d)$ if $a \preceq b$, $c \succeq d$, and the interval of classes for a, b is included in the interval of classes for c, d.

The second motivation is related to the choice of the preference model type. We chose the logical preference model in terms of monotonic decision rules. This is because axiomatic analysis of all three preference model types leads to the conclusion that decision rules, as they are defined in DRSA, are the only aggregation operators that give account of most complex interactions among criteria, are non-compensatory, accept ordinal evaluation scales and do not convert ordinal evaluations into cardinal ones [14]. Rules identify values that drive DM's decisions—each rule is a scenario of a causal relationship between evaluations on a subset of criteria and a comprehensive judgment. They are also easily interpretable by users who trust proposed recommendations more [13].

In this introduction, we should also refer to ranking methods based on preference learning in a way proposed by Machine Learning (ML) [10]. In ML, this task is known as "learning to rank" and also involves learning of a preference model from pairwise comparisons of some alternatives (called items in ML) [5, 9, 18]. Precisely, the pairwise comparisons are provided by users (DMs) as lists of items with some partial order between items in each list. This information is called the training data. Machine preference learning consists in discovering a model that predicts preference for a new set of items (or the input set of items considered in a different context) so that the produced ranking is statistically "similar" to the order provided as the training data. In this approach, learning is traditionally achieved by minimizing an empirical estimate of an assumed loss function on rankings [6]. Learning to rank emerged to address application needs in areas such as information retrieval, Internet-related applications, and bio-informatics. Indeed, ranking is at the core of document retrieval, collaborative filtering, or computational advertising. In recommender systems, a ranked list of related items should be recommended to a user who has shown interest in some other items. In computational biology, one ranks candidate structures in protein structure prediction problem, whereas in proteomics there is a need for the identification of frequent top scoring peptides.

MCDA and machine preference learning show many similarities, however, there are also striking differences between them [4]. In particular, MCDA stimulates the DM to interact with the method by incrementally enriching the preference information and observing its consequences on the recommended rankings. This feature reveals a specific aspect of learning adopted in MCDA which contrasts with ML oriented towards preference discovery without interaction with the DM. For the purpose of this article, we should also stress that neither in MCDA nor in machine preference learning, the ordinal intensity of preference has been considered in the way explained above as the first motivation.

The paper is organized as follows. In Sect. 2, we define notation and basic concepts related to the preference information and approximated sets of pairs of alternatives. Induction of decision rules from rough approximations is described in Sect. 3. Application of decision rules on a set of alternatives is a subject of Sect. 4. In Sect. 5, exploitation of a preference graph resulting from application of decision rules is presented together with the end result in form of a weak order on the set of alternatives. In Sect. 6, an illustrative example shows how the proposed method can be applied in a hypothetical case study. The final section includes summary and conclusions.

2 Notation and Basic Concepts

We shall use the following notation:

- $A = \{x, y, \ldots\}$—a finite set of alternatives to be ranked;
- $C_1, C_2, \ldots, C_p$—p predefined preference-ordered classes, where C_{h+1} is preferred to C_h, $h = 1, \ldots, p-1$; moreover, $H = \{1, \ldots, p\}$ denotes the set of class indices;
- $A^R = \{a, b, \ldots\}$—a finite set of m reference alternatives, on which the DM accepts to express holistic preferences, such that each reference alternative is assigned to one of the classes $C_1, C_2, \ldots, C_p$; we assume that $A^R \subseteq A$;
- $B = A^R \times A^R$;
- $G = \{g_1, g_2, \ldots, g_j, \ldots, g_n\}$—a finite set of n evaluation criteria with ordinal or cardinal scales; without loss of generality, we assume that all criteria are of gain type, i.e., the greater the criterion value, the better.

A criterion with the cardinal scale is called a cardinal criterion; the set of all cardinal criteria is denoted by $G^N \subseteq G$, while $\mathcal{J}_{G^N}$ denotes their indices. A criterion with the ordinal scale is called an ordinal criterion; the set of all ordinal criteria is denoted by $G^O \subseteq G$, while $\mathcal{J}_{G^O}$ denotes their indices. Moreover, $G^N \cup G^O = G$ and $G_N \cap G_O = \emptyset$. For the sake of simplicity, we assume that for each cardinal criterion $g_j \in G^N$, intensity of preference of a over b is defined as the difference of evaluations, i.e., it is equal to $\Delta_j(a,b) = g_j(a) - g_j(b)$. In case of criterion $g_j \in G^O$ with an ordinal scale, one can only establish an order of evaluations $g_j(a)$, $a \in A$.

Preference information. We assume that the DM provides a set of assignment examples, each one consisting of a reference alternative $a \in A^R$ and its assignment $Cl_{DM}(a) = C_i$, $1 \leq i \leq p$.

Dominance relation for pairs of alternatives. The pair of alternatives $(a,b) \in B$ *dominates* pair $(c,d) \in B$ with respect to set of criteria G, denoted by $(a,b)D_2(c,d)$, if and only if (iff):

- for all $g_j \in G^N$, $\Delta_j(a,b) \geq \Delta_j(c,d)$, where $\Delta_j(a,b) = g_j(a) - g_j(b)$;
- for all $g_j \in G^O$, $g_j(a) \geq g_j(c)$ and $g_j(b) \leq g_j(d)$.

Dominance relation D_2 is a partial weak order on B. If $(a,b)D_2(c,d)$, one expects that not $Cl_{DM}(c) \geq Cl_{DM}(a)$ and $Cl_{DM}(d) \leq Cl_{DM}(b)$, with at least one of these relations being strict. Violation of this principle is considered as an inconsistency with respect to dominance relation D_2 on B and the order imposed on considered classes.

Granules of knowledge. The set of pairs of alternatives dominating $(a,b) \in B$, $D_2^+(a,b)$, is called the *dominating set* or *positive dominance cone*:

$$D_2^+(a,b) = \{(c,d) \in B : (c,d)D_2(a,b)\}. \tag{1}$$

The set of pairs of alternatives dominated by (a,b), $D_2^-(a,b)$, is called the *dominated set* or *negative dominance cone*:

$$D_2^-(a,b) = \{(c,d) \in B : (a,b)D_2(c,d)\}. \tag{2}$$

Approximated sets of pairs of alternatives. Since classes $C_1, \ldots, C_p$ are preference-ordered, when comparing two reference alternatives $a, b \in A^R$, where $a \in C_i$ and $b \in C_j$, $1 \leq i, j \leq p$, three possibilities may arise:

- $i = j$, which means that a is indiscernible with b,
- $i > j$, which means that a is preferred to b,
- $i < j$, which means that b is preferred to a.

Notice that the comparison of two reference alternatives has an ordinal character only. In consequence, given two pairs of reference alternatives $(a,b), (c,d) \in B$, one can compare them with respect to preference only if the interval of classes to which a, b belong includes (or is included in) the interval of classes to which c, d belong. Specifically, if $Cl_{DM}(a) \geq Cl_{DM}(c)$ and $Cl_{DM}(b) \leq Cl_{DM}(d)$, then a is preferred to b at least as much as c is preferred to d (c is preferred to d at most as much as a is preferred to b). Otherwise, these pairs are incomparable.

Let us consider the following set of pairs of alternatives

$$S^{i,j} = \{(a,b) \in B : Cl_{DM}(a) = C_i \text{ and } Cl_{DM}(b) = C_j\}, \tag{3}$$

where $1 \leq i, j \leq p$. It includes pairs of alternatives with the same preference of the first alternative over the second one.

Using definition (3), one can define the following unions of sets:

$$S^{\succeq(i,j)} = \bigcup_{k \geq i, l \leq j} S^{k,l}; \tag{4}$$

$$S^{\preceq(i,j)} = \bigcup_{k \leq i, l \geq j} S^{k,l}, \tag{5}$$

where $1 \leq i, j, k, l \leq p$. Union $S^{\succeq(i,j)}$ is a set composed of pairs of reference alternatives (a, b) such that a is preferred to b at least as much as c is preferred to d, where $(c, d) \in S^{i,j}$. Analogously, union $S^{\preceq(i,j)}$ is a set composed of pairs of reference alternatives (a, b) such that a is preferred to b at most as much as c is preferred to d, where $(c, d) \in S^{i,j}$.

Notice that the above unions are binary relations, thus, the expressions $(a, b) \in S^{\succeq(i,j)}$ and $aS^{\succeq(i,j)}b$ can be used alternatively. Moreover, the unions are related in the following way: $S^{\succeq(i,j)} \supseteq S^{\succeq(k,l)}$ if and only if $i \leq k$ and $j \geq l$, for all $i, j, k, l \in H$; analogously, $S^{\preceq(i,j)} \supseteq S^{\preceq(k,l)}$ if and only if $i \geq k$ and $j \leq l$, for all $i, j, k, l \in H$. Thus, there exist two lattices of unions $S^{\succeq(i,j)}$ and $S^{\preceq(i,j)}$, respectively, ordered by weak inclusion relation. These lattices can be depicted by Hasse diagrams.

Figure 1 presents Hasse diagram of lattice of unions $S^{\succeq(i,j)}$, while Fig. 2 presents Hasse diagram of lattice of unions $S^{\preceq(i,j)}$, $1 \leq i, j \leq 4$; in both diagrams arcs show the direction of inclusion, i.e., an arc leading from union U_1 to U_2 marks inclusion $U_1 \subseteq U_2$. For example, in case of union $S^{\succeq(3,2)}$:

$$S^{\succeq(3,2)} \supseteq S^{\succeq(3,1)} \supseteq S^{\succeq(4,1)} \text{ and}$$
$$S^{\succeq(3,2)} \supseteq S^{\succeq(4,2)} \supseteq S^{\succeq(4,1)}.$$

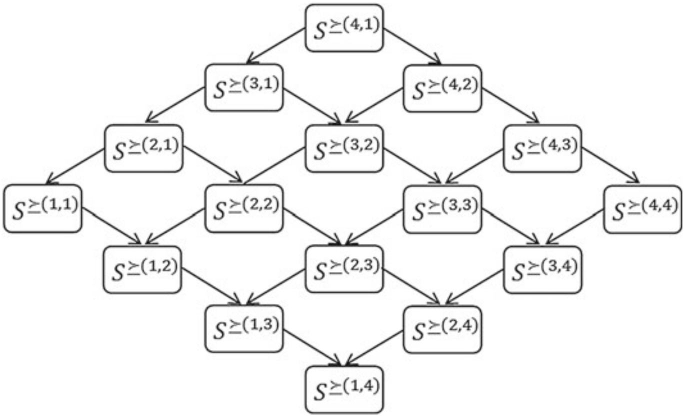


Fig. 1 Hasse diagram of the lattice of unions $S^{\succeq(i,j)}$ for $1 \leq i, j \leq 4$

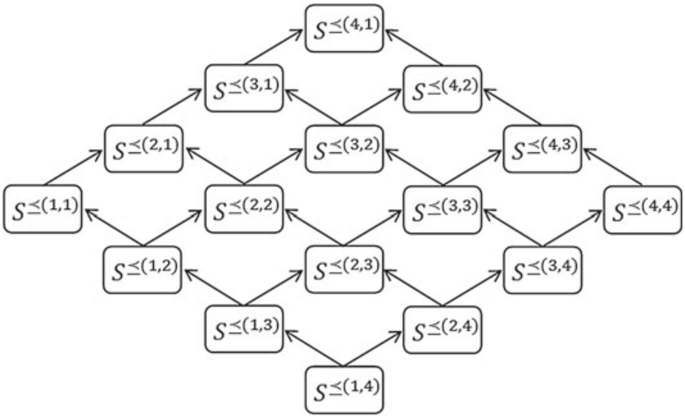


Fig. 2 Hasse diagram of the lattice of unions $S^{\preceq(i,j)}$ for $1 \leq i, j \leq 4$

Moreover, in case of union $S^{\preceq(2,2)}$:

$$S^{\preceq(2,2)} \supseteq S^{\preceq(1,2)} \supseteq S^{\preceq(1,3)} \supseteq S^{\preceq(1,4)} \text{ and}$$
$$S^{\preceq(2,2)} \supseteq S^{\preceq(2,3)} \supseteq S^{\preceq(1,3)} \supseteq S^{\preceq(1,4)} \text{ and}$$
$$S^{\preceq(2,2)} \supseteq S^{\preceq(2,3)} \supseteq S^{\preceq(2,4)} \supseteq S^{\preceq(1,4)}.$$

Rough approximations. We approximate unions $S^{\succeq(i,j)}$ using positive dominance cones $D_2^+(\cdot,\cdot)$, and unions $S^{\preceq(i,j)}$ using negative dominance cones $D_2^-(\cdot,\cdot)$. The lower and upper approximations of $S^{\succeq(i,j)}$ and $S^{\preceq(i,j)}$ are defined, respectively, as:

$$\underline{S^{\succeq(i,j)}} = \{(a,b) \in S^{\succeq(i,j)} : D_2^+(a,b) \cap (S^{\preceq(i-1,j)} \cup S^{\preceq(i,j+1)}) = \emptyset\}; \quad (6)$$
$$\overline{S^{\succeq(i,j)}} = \{(a,b) \in B : D_2^-(a,b) \cap S^{\succeq(i,j)} \neq \emptyset\}; \quad (7)$$
$$\underline{S^{\preceq(i,j)}} = \{(a,b) \in S^{\preceq(i,j)} : D_2^-(a,b) \cap (S^{\succeq(i,j-1)} \cup S^{\succeq(i+1,j)}) = \emptyset\}; \quad (8)$$
$$\overline{S^{\preceq(i,j)}} = \{(a,b) \in B : D_2^+(a,b) \cap S^{\preceq(i,j)} \neq \emptyset\}, \quad (9)$$

where $1 \leq i, j \leq p$.

Finally, the boundaries of $S^{\succeq(i,j)}$ and $S^{\preceq(i,j)}$ are defined, respectively, as:

$$Bn(S^{\succeq(i,j)}) = \overline{S^{\succeq(i,j)}} \setminus \underline{S^{\succeq(i,j)}}, \quad (10)$$
$$Bn(S^{\preceq(i,j)}) = \overline{S^{\preceq(i,j)}} \setminus \underline{S^{\preceq(i,j)}}. \quad (11)$$

Let us explain the idea underlying definitions of $\underline{S^{\succeq(i,j)}}$ and $\overline{S^{\succeq(i,j)}}$. On one hand, $\underline{S^{\succeq(i,j)}}$ contains pairs of reference alternatives $(a,b) \in B$ which are not dominated by any pair of reference alternatives $(c,d) \in B$ such that the class of c is worse (worse or equal) than that of a and the class of d is better or equal (better) than that of b. For example, the lower approximation of $S^{\succeq(3,2)}$ contains pairs of reference alternatives (a, b) which are not dominated by any pair (c, d) belonging to $S^{\preceq(2,2)}$ or $S^{\preceq(3,3)}$. On the other hand, $\overline{S^{\succeq(i,j)}}$ contains pairs of reference alternatives $(a, b) \in B$ which dominate at least one pair of reference alternatives (c, d) belonging to $S^{\succeq(i,j)}$.

3 Induction of Decision Rules

We assume that a preference model of the DM is a set of minimal decision rules, being statements of the type: "*if premise, then conclusion*" that represent a form of dependency between the condition part and the decision part. The premise of a rule is a conjunction of elementary conditions concerning individual criteria, and the decision part of a rule suggests that a pair of alternatives covered by the rule should be assigned to particular union $S^{\succeq(i,j)}$ or $S^{\preceq(i,j)}$, $1 \leq i, j \leq p$. We say that a pair of alternatives is *covered* by a decision rule if it matches the premise of the rule. On the other hand, a pair of alternatives *supports* a decision rule if it matches both premise and conclusion of the rule. Although we can distinguish certain, possible, and approximate rules, in this paper we will focus on the certain rules only.

In order to induce certain decision rules with conclusion $xS^{\succeq(i,j)}y$ or $xS^{\preceq(i,j)}y$, $1 \leq i, j \leq p$, one needs to consider:

- *positive examples*, i.e., consistent pairs of reference alternatives concordant with given conclusion (pairs of reference alternatives from the lower approximation $\underline{S^{\succeq(i,j)}}$ or $\underline{S^{\preceq(i,j)}}$, respectively), and
- *negative examples*, i.e., pairs of reference alternatives contained in $S^{\preceq(i-1,j)} \cup S^{\preceq(i,j+1)}$ or $S^{\succeq(i,j-1)} \cup S^{\succeq(i+1,j)}$, respectively.

Observe that sets of positive and negative examples do not make partition of B. Apart from both types of examples, B includes also so-called *neutral examples*, i.e., pairs of reference alternatives that belong to $B \setminus (S^{\succeq(i,j)} \cup S^{\preceq(i-1,j)} \cup S^{\preceq(i,j+1)})$ or $B \setminus (S^{\preceq(i,j)} \cup S^{\succeq(i,j-1)} \cup S^{\succeq(i+1,j)})$, respectively. These examples are not taken into account during rule induction.

In the following, when defining the syntax of decision rules, instead of concise conclusion $xS^{\succeq(i,j)}y$, we employ the equivalent expression $Cl_{DM}(x) \geq C_i$ and $Cl_{DM}(y) \leq C_j$, which is more informative for the DM. For the same reason, instead of conclusion $xS^{\preceq(i,j)}y$, we use the expression $Cl_{DM}(x) \leq C_i$ and $Cl_{DM}(y) \geq C_j$.

We distinguish two types of decision rules:

- "at least" decision rules, with the following syntax:

$$\begin{aligned}&\text{if } \Delta_{j_1}(x, y) \geq \delta_{j_1} \text{ and } \ldots \text{ and } \Delta_{j_v}(x, y) \geq \delta_{j_v} \text{ and } \ldots \text{ and } g_{j_{v+1}}(x) \geq r_{j_{v+1}}\\ &\qquad \text{and } g_{j_{v+1}}(y) \leq s_{j_{v+1}} \text{ and } \ldots \text{ and } g_{j_z}(x) \geq r_{j_z} \text{ and } g_{j_z}(y) \leq s_{j_z}\\ &\qquad\qquad \text{then } Cl_{DM}(x) \geq C_i \text{ and } Cl_{DM}(y) \leq C_j,\end{aligned}$$

- "at most" decision rules, with the following syntax:

$$\text{if } \Delta_{j_1}(x,y) \leq \delta_{j_1} \text{ and } \dots \text{ and } \Delta_{j_v}(x,y) \leq \delta_{j_v} \text{ and } \dots \text{ and } g_{j_{v+1}}(x) \leq r_{j_{v+1}}$$
$$\text{and } g_{j_{v+1}}(y) \geq s_{j_{v+1}} \text{ and } \dots \text{ and } g_{j_z}(x) \leq r_{j_z} \text{ and } g_{j_z}(y) \geq s_{j_z}$$
$$\text{then } Cl_{DM}(x) \leq C_i \text{ and } Cl_{DM}(y) \geq C_j,$$

where $\delta_{j_k} \in \{g_{j_k}(a) - g_{j_k}(b) : (a,b) \in B\} \subseteq \Re$, for $j_k \in \{j_1, \dots, j_v\} \subseteq \mathcal{J}_{G^N}$; $(r_{j_k}, s_{j_k}) \in \{(g_{j_k}(a), g_{j_k}(b)) : (a,b) \in B\} \subseteq \Re \times \Re$, for $j_k \in \{j_{v+1}, \dots, j_z\} \subseteq \mathcal{J}_{G^O}$. For instance, considering ranking of cars, a decision rule could be "if car x has maximum speed at least 25 km/h greater than car y (cardinal criterion), and car x has comfort at least 3 while car y has comfort at most 2 (ordinal criterion), then car x is assigned to class at least C_3 while car y is assigned to class at most C_1", where values 2 and 3 code ordinal evaluations 'medium' and 'good', respectively.

The sets of "at least" and "at most" decision rules with the above syntax can be induced using one of the well-known rule induction algorithms, e.g., VC-DomLEM algorithm [2, 3], DomLEM algorithm [15, 23] or LEM2 algorithm [16, 17].

4 Application of Decision Rules

After induction of decision rules, the next step of the proposed methodology for multiple criteria ranking is the application of induced rules on set A. This application yields a preference structure on set A. Each pair of alternatives $(x, y) \in A \times A$ can be covered by some decision rules suggesting assignment to relations $S^{\succeq(i,j)}$ and/or to relations $S^{\preceq(i,j)}$, $1 \leq i, j \leq p$. It can be also not covered by any rule. In order to represent these situations, we introduce the following binary relations on A:

$$\mathbb{S}^{\succeq(i,j)} = \{(x,y) \in A \times A : \exists r \in R_{S^{\succeq(i,j)}} \text{ such that } r \text{ covers } (x,y)\}, \quad (12)$$
$$\mathbb{S}^{\preceq(i,j)} = \{(x,y) \in A \times A : \exists r \in R_{S^{\preceq(i,j)}} \text{ such that } r \text{ covers } (x,y)\}, \quad (13)$$

where $1 \leq i, j \leq p$ and $R_{S^{\succeq(i,j)}}$ denotes set of rules with conclusion $Cl_{DM}(x) \geq C_i$ and $Cl_{DM}(y) \leq C_j$. Notice that $S^{\succeq(i,j)}$ and $\mathbb{S}^{\succeq(i,j)}$ are two different relations. The first one (see Definition (4)) is defined on set A^R and concerns class assignments of reference alternatives, while the second one, introduced above, is defined on set A and concerns coverage by induced decision rules.

The preference structure on A, composed of relations $\mathbb{S}^{\succeq(i,j)}$ and $\mathbb{S}^{\preceq(i,j)}$, $1 \leq i, j \leq p$, can be represented by a *preference graph*. It is a directed multigraph $\mathcal{G}$. Each vertex (node) v_x of the preference graph corresponds to exactly one alternative $x \in A$. One can distinguish in $\mathcal{G}$ two types of arcs: $\mathbb{S}^{\succeq(i,j)}$-arcs and $\mathbb{S}^{\preceq(i,j)}$-arcs. For example, given $p = 4$, the preference graph features an $\mathbb{S}^{\succeq(4,1)}$-arc from vertex v_x to v_y iff $x\mathbb{S}^{\succeq(4,1)}y$. $\mathcal{G}$ is a multigraph since for any pair of alternatives $(x, y) \in A \times A$ there may be more than one arc from vertex v_x to vertex v_y. A *final recommendation* for the multiple criteria ranking problem at hand, in terms of a complete weak order of all alternatives belonging to set A, can be obtained upon a suitable *exploitation* of the preference graph.

5 Exploitation of the Preference Graph

For the purpose of exploitation of the preference graph we employ the Weighted Fuzzy Net Flow Score (WFNFS) procedure described in [8]. As proved in [8], this procedure ensures that the obtained ranking contains dominance relation on set A, i.e., if alternative a dominates b with respect to set of criteria G, then a is going to be ranked not lower than b. Since relations given by (12) and (13) are crisp, in the following we describe a simplified version of this procedure that we call Weighted Net Flow Score (WNFS) procedure.

Let [] denote the Iverson bracket function defined as:

$$[P] = \begin{cases} 1 \text{ if } P \text{ is true,} \\ 0 \text{ otherwise.} \end{cases} \tag{14}$$

In order to exploit a preference graph resulting from application of decision rules on set A, we employ *scoring function* $NFS : A \rightarrow \Re$ defined as

$$\begin{aligned} NFS(x) = \sum_{y \in A \setminus \{x\}} \Bigg(& \sum_{1 \leq j \leq i \leq p} w^{\succeq(i,j)} \big([x \mathbb{S}^{\succeq(i,j)} y] - [y \mathbb{S}^{\succeq(i,j)} x] \big) \\ & - \sum_{1 \leq i \leq j \leq p} w^{\preceq(i,j)} \big([x \mathbb{S}^{\preceq(i,j)} y] - [y \mathbb{S}^{\preceq(i,j)} x] \big) \Bigg), \end{aligned} \tag{15}$$

where weights $w^{\succeq(i,j)}$, for $i \geq j$, and weights $w^{\preceq(i,j)}$, for $i \leq j$, are by default equal to one but can be set different by the DM, e.g., in order to express greater importance of preference between alternatives from more distant classes. For each alternative $x \in A$, $NFS(x)$ takes into account two types of arguments in favor of x (i.e., existence of $y \in A \setminus \{x\}$ such that $x\mathbb{S}^{\succeq(i,j)}y$, and existence of $y \in A \setminus \{x\}$ such that $y\mathbb{S}^{\preceq(i,j)}x$) and two types of arguments in disfavor of x (i.e., existence of $y \in A \setminus \{x\}$ such that $y\mathbb{S}^{\succeq(i,j)}x$, and existence of $y \in A \setminus \{x\}$ such that $x\mathbb{S}^{\preceq(i,j)}y$). In the following, we will also use the notions of *strength* and *weakness* of an alternative, with respect to $\mathbb{S}^{\succeq(i,j)}$ and $\mathbb{S}^{\preceq(i,j)}$. For instance, the strength of $x \in A$ with respect to $\mathbb{S}^{\succeq(i,j)}$ is the value $\sum_{y \in A \setminus \{x\}} \sum_{1 \leq j \leq i \leq p} w^{\succeq(i,j)}[x\mathbb{S}^{\succeq(i,j)}y]$, while the weakness of $x \in A$ with respect to $\mathbb{S}^{\preceq(i,j)}$ is the value $\sum_{y \in A \setminus \{x\}} \sum_{1 \leq i \leq j \leq p} w^{\preceq(i,j)}[x\mathbb{S}^{\preceq(i,j)}y]$. Notice that if we put weights $w^{\succeq(i,j)}$ and $w^{\preceq(i,j)}$ on respective arcs of the considered preference graph, then calculation of $NFS(x)$ is equivalent to the calculation of a kind of net flow of vertex v_x of this graph, where positive and negative inflows and outflows are considered. Function *NFS* induces a weak order on A, which is a solution of the considered multiple criteria ranking problem.

6 Illustrative Example

In this section, we illustrate the use of the presented method by considering real-world data concerning 13 Polish research units (group of joint evaluation, called SI3MU). The units are evaluated on the following three gain-type criteria:

- scientific activity (g_1) including scientific publications in international journals and number of patents; the evaluation reflects an average number of points gained by a single researcher of the unit;
- scientific potential (g_2) including the ability to grant scientific degrees, number of professor titles obtained by researchers of this unit in the evaluation period, as well as prestigious memberships of the researchers; all achievements are scored and these scores are summed up to get an evaluation;
- material effects of unit's activities (g_3) representing money acquired from grants or cooperation with industry.

The performances of the 13 considered research units (denoted by a to m) are given in Table 1. The objective of the study is to order the units from the best to the worst. Although the aim consists in delivering a ranking, we will employ preference information which is specific for multiple criteria sorting problems.

Assignment examples. The preference information consists of exemplary class assignments for 6 randomly chosen reference units (these are distinguished with a non-empty entry in Table 1 (column $Ref.$)). There are 2 units assigned to each of the three considered classes Cl_1–Cl_3 (with Cl_3 being the best class). The assignment examples are derived from the original classification provided by the Polish Ministry of Science and Higher Education in 2014. Our aim is to "learn" a rule preference model on the 6 assignment examples, and apply this model on the whole set of 13 units in order to rank them.

Rough approximations. The provided 6 assignment examples entail consideration of 36 pairs of reference alternatives. It turns out that there is no inconsistency with respect to dominance relation D_2 on the set of these pairs and the order imposed on considered classes. The lower approximations for selected unions $S^{\succeq(i,j)}$ and $S^{\preceq(i,j)}$, $1 \leq i, j \leq 3$, are presented in Table 2. Obviously, in this case they are equal to upper approximations.

Minimal sets of decision rules. To induce decision rules from the lower approximations given in Table 2, we used a heuristic algorithm of a sequential covering type (inspired by LEM2). When selecting conditions for inclusion in a decision rule, we preferred these conditions that allowed to cover maximal number of positive examples, and then, to break possible ties, conditions which allowed to cover minimal number of negative examples. The resulting minimal sets of decision rules for selected unions are listed in Table 3. Each of them consists of just a single decision rule with at most two conditions.

Table 1 Research units' performances and class assignments for reference units (column *Ref.*)

Unit	g_1	g_2	g_3	*Ref.*
a	29.41	127	270.93	Cl_3
b	30.57	122	280.14	–
c	31.34	283	122.78	Cl_3
d	46.46	117	34.44	Cl_2
e	15.99	50	155.55	–
f	20.08	108	47.43	Cl_2
g	17.03	60	61.12	–
h	23.65	150	27.22	–
i	9.54	109	50.65	–
j	11.41	106	28.39	Cl_1
k	10.98	2	13.28	Cl_1
l	9.66	0	11.69	–
m	4.16	2	1.35	–

Table 2 Lower approximations for selected unions $S^{\succeq(i,j)}$, $S^{\preceq(i,j)}$, $1 \leq i, j \leq 3$

$\underline{S^{\succeq(3,1)}}$	=	$\{(a, j), (a, k), (c, j), (c, k)\}$
$\underline{S^{\succeq(3,2)}}$	=	$\{(a, j), (a, k), (c, j), (c, k), (a, d), (a, f), (c, d), (c, f)\}$
$\underline{S^{\succeq(2,1)}}$	=	$\{(a, j), (a, k), (c, j), (c, k), (d, j), (d, k), (f, j), (f, k)\}$
$\underline{S^{\succeq(3,3)}}$	=	$\{(a, j), (a, k), (c, j), (c, k), (a, d), (a, f), (c, d), (c, f), (a, a), (a, c), (c, a), (c, c)\}$
$\underline{S^{\succeq(2,2)}}$	=	$\{(a, j), (a, k), (c, j), (c, k), (a, d),(a, f), (c, d),(c, f), (d, j),(d, k), (f, j), (f, k),$ $(d, d), (d, f), (f, d), (f, f)\}$
$\underline{S^{\succeq(1,1)}}$	=	$\{(a, j), (a, k), (c, j), (c, k), (d, j), (d, k), (f, j), (f, k), (j, j), (j, k), (k, j), (k, k)\}$
$\underline{S^{\preceq(1,3})}$	=	$\{(j, a), (k, a), (j, c), (k, c)\}$
$\underline{S^{\preceq(2,3)}}$	=	$\{(j, a), (k, a), (j, c), (k, c), (d, a), (f, a), (d, c), (f, c)\}$
$\underline{S^{\preceq(1,2)}}$	=	$\{(j, a), (k, a), (j, c), (k, c), (j, d), (k, d), (j, f), (k, f)\}$
$\underline{S^{\preceq(3,3)}}$	=	$\{(j, a), (k, a), (j, c), (k, c), (d, a), (f, a), (d, c), (f, c)(a, a), (a, c), (c, a), (c, c)\}$
$\underline{S^{\preceq(2,2)}}$	=	$\{(j, a), (k, a), (j, c), (k, c), (d, a), (f, a), (d, c), (f, c), (j, d), (k, d), (j, f), (k, f),$ $(d, d), (d, f), (f, d), (f, f)\}$
$\underline{S^{\preceq(1,1)}}$	=	$\{(j, a), (k, a), (j, c), (k, c), (j, d), (k, d), (j, f), (k, f), (j, j), (j, k), (k, j), (k, k)\}$

Weights of arcs in the preference graph. To compute the score $NFS(x)$ of each unit $x \in A$, we employed the weights $w^{\succeq(i,j)}$ and $w^{\preceq(i,j)}$ given in Table 4. These weights were chosen to be symmetric, i.e., $w^{\succeq(i,j)} = w^{\preceq(j,i)}$. Moreover, they were set so that to express:

- greater importance of preference between units from more distant classes, e.g., $w^{\succeq(3,1)} = 6 > w^{\succeq(3,2)} = 3 > w^{\succeq(3,3)} = 1$;

Table 3 Minimal sets of decision rules induced from lower approximations of selected unions $S^{\succeq(i,j)}$, $S^{\preceq(i,j)}$, $1 \leq i, j \leq 3$

Appr.	Rule
$\underline{S}^{\succeq(3,1)}$	if $\Delta_1(x, y) \geq 18.0$ and $\Delta_3(x, y) \geq 94.39$ then $Cl(x) \geq C_3$ and $Cl(y) \leq C_1$
$\underline{S}^{\succeq(3,2)}$	if $\Delta_2(x, y) \geq 10.0$ and $\Delta_3(x, y) \geq 75.35$ then $Cl(x) \geq C_3$ and $Cl(y) \leq C_2$
$\underline{S}^{\succeq(2,1)}$	if $\Delta_1(x, y) \geq 8.67$ and $\Delta_3(x, y) \geq 6.05$ then $Cl(x) \geq C_2$ and $Cl(y) \leq C_1$
$\underline{S}^{\succeq(3,3)}$	if $\Delta_1(x, y) \geq -156.0$ and $\Delta_3(x, y) \geq -148.15$ then $Cl(x) \geq C_3$ and $Cl(y) \leq C_3$
$\underline{S}^{\succeq(2,2)}$	if $\Delta_1(x, y) \geq -12.99$ and $\Delta_3(x, y) \geq -9.0$ then $Cl(x) \geq C_2$ and $Cl(y) \leq C_2$
$\underline{S}^{\succeq(1,1)}$	if $\Delta_1(x, y) \geq -0.43$ then $Cl(x) \geq C_1$ and $Cl(y) \leq C_1$
$\underline{S}^{\preceq(1,3)}$	if $\Delta_1(x, y) \leq -18.0$ and $\Delta_3(x, y) \leq -94.39$ then $Cl(x) \leq C_1$ and $Cl(y) \geq C_3$
$\underline{S}^{\preceq(2,3)}$	if $\Delta_2(x, y) \leq -10.0$ and $\Delta_3(x, y) \leq -75.35$ then $Cl(x) \leq C_2$ and $Cl(y) \geq C_3$
$\underline{S}^{\preceq(1,2)}$	if $\Delta_1(x, y) \leq -8.67$ and $\Delta_3(x, y) \leq -6.05$ then $Cl(x) \leq C_1$ and $Cl(y) \geq C_2$
$\underline{S}^{\preceq(3,3)}$	if $\Delta_1(x, y) \leq 156.0$ and $\Delta_3(x, y) \leq 148.15$ then $Cl(x) \leq C_3$ and $Cl(y) \geq C_3$
$\underline{S}^{\preceq(2,2)}$	if $\Delta_1(x, y) \leq 12.99$ and $\Delta_3(x, y) \leq 9.0$ then $Cl(x) \leq C_2$ and $Cl(y) \geq C_2$
$\underline{S}^{\preceq(1,1)}$	if $\Delta_1(x, y) \leq 0.43$ then $Cl(x) \leq C_1$ and $Cl(y) \geq C_1$

Table 4 Weights of arcs in the preference graph

$w^{\succeq(3,1)}$	$w^{\succeq(3,2)}$	$w^{\succeq(2,1)}$	$w^{\succeq(1,1)}$	$w^{\succeq(2,2)}$	$w^{\succeq(3,3)}$
6	3	3	1	1	1
$w^{\preceq(1,3)}$	$w^{\preceq(2,3)}$	$w^{\preceq(1,2)}$	$w^{\preceq(1,1)}$	$w^{\preceq(2,2)}$	$w^{\preceq(3,3)}$
6	3	3	1	1	1

- equal importance of preference for pairs of units for which the class difference is the same, e.g., $w^{\succeq(3,2)} = w^{\succeq(2,1)} = 3$.

Ranking. In Table 5, we show the result of application of the induced decision rules on set A, and aggregation of the resulting relations $\mathbb{S}^{\succeq(i,j)}$ and $\mathbb{S}^{\preceq(i,j)}$ using scoring function NFS with the weights provided in Table 4. The obtained ranking is unique for the induced set of decision rules and adopted weights. For clarity, for each research unit we additionally decompose its comprehensive score into the strength and weakness derived from both $\mathbb{S}^{\succeq(i,j)}$ and $\mathbb{S}^{\preceq(i,j)}$. The final ranking (see Table 5, column $Rank(x)$) reproduces the preference order derived from assignments examples, i.e., all reference units assigned to class Cl_{h+1} are ranked better than these assigned to class Cl_h, for $h = 1, 2$. Moreover, when compared with the original classification of the Polish Ministry of Science and Higher Education, all units judged as the best (worst) ones by the Ministry, i.e., a to c (j to m), attain clearly positive (negative) scores in our procedure.

Table 5 Final scores (column $NFS(x)$) and ranks (column $Rank(x)$) of research units (for each unit, we provide its strength and weakness reflected in $NFS(x)$, both with respect to $\mathbb{S}^{\succeq(i,j)}$ and $\mathbb{S}^{\preceq(i,j)}$)

Unit (x)	$\mathbb{S}^{\succeq(i,j)}$		$\mathbb{S}^{\preceq(i,j)}$		$NFS(x)$	$Rank(x)$	$Ref.$
	Strength	Weakness	Strength	Weakness			
a	116	10	117	10	213	1	Cl_3
b	113	9	114	8	210	2	–
c	100	8	100	9	183	3	Cl_3
d	41	28	45	30	28	4	Cl_2
e	39	37	41	35	8	5	–
f	37	46	38	45	−16	7	Cl_2
g	27	42	29	41	−27	8	–
h	35	31	35	34	5	6	–
i	26	54	22	59	−63	9	–
j	23	72	21	75	−103	10	Cl_1
k	16	85	18	84	−135	11	Cl_1
l	15	86	16	86	−141	12	–
m	13	93	13	93	−160	13	–

7 Summary and Conclusions

We presented a new method for multiple criteria ranking problem, characterized by the following features:

- the preference information provided by the DM has the form of sorting examples, i.e., assignments of some reference alternatives to pre-defined and ordered quality classes,
- the intensity of preference between any two alternatives is considered as purely ordinal, i.e., the number of quality classes separating two assigned alternatives is not meaningful for intensity of preference,
- the intensity of preference for pairs of quality classes can be represented by a lattice depicted by Hasse diagram, i.e., one can say that intensity of preference for a pair of alternatives is greater than that of another pair, only if the interval of classes for the first pair includes that of the second pair,
- the method employs the decision rule preference model—the rules are induced from rough approximations of unions of preference intensity relations, without converting the ordinal input preference information into cardinal one,
- the set of rules is an easy to read summary of scenarios of causal relationships between evaluations of pairs of reference alternatives on a subset of criteria and a comprehensive judgment,

- application of decision rules on a considered set of alternatives leads to a preference graph—its exploitation using the weighted net flow score procedure results in a linear ranking.

In conclusion, one can observe that the proposed method does what was promised: starting from an ordinal preference information about intensity of preference on a subset of alternatives, it builds an intelligible preference model being compatible with the input preference information, and applies this model on the whole set of considered alternatives to finally rank them from the best to the worst. An illustrative case study performed at the end of this paper supports this claim.

Acknowledgments The first author acknowledges financial support from the National Science Center (grant no. DEC-2013/11/D/ST6/03056). The third author declares that he is a scholarship holder within the 2012/2013 project "Scholarship support for Ph.D. students specializing in majors strategic for Wielkopolska's development", Sub-measure 8.2.2 of Human Capital Operational Programme, co-financed by European Union under the European Social Fund.

References

1. Bana e Costa CA, Vansnick J-C (1994) MACBETH: an interactive path towards the construction of cardinal value functions. Int Trans Oper Res 1(4):387–500
2. Błaszczyński J, Słowiński R, Szeląg M (2010) Probabilistic rough set approaches to ordinal classification with monotonicity constraints. In: Hüllermeier E, Kruse R, Hoffmann F (eds) IPMU 2010. Lecture notes in artificial intelligence, vol 6178. Springer, Berlin, pp 99–108
3. Błaszczyński J, Słowiński R, Szeląg M (2011) Sequential covering rule induction algorithm for variable consistency rough set approaches. Inf Sci 181:987–1002
4. Corrente S, Greco S, Kadziński M, Słowiński R (2013) Robust ordinal regression in preference learning and ranking. Mach Learn 93:381–422
5. Dembczyński K, Kotłowski W, Słowiński R, Szeląg M (2010) Learning of rule ensembles for multiple attribute ranking problems. In: Fürnkranz J, Hüllermeier E (eds) Preference learning. Springer, Berlin, pp 217–247
6. Doumpos M, Zopounidis C (2012) Preference disaggregation and statistical learning for multicriteria decision support: a review. Eur J Oper Res 209(3):203–214
7. Figueira J, Greco S, Słowiński R (2009) Building a set of additive value functions representing a reference preorder and intensities of preference: grip method. Eur J Oper Res 195(2):460–486
8. Fortemps P, Greco S, Słowiński R (2008) Multicriteria decision support using rules that represent rough-graded preference relations. Eur J Oper Res 188(1):206–223
9. Fürnkranz J, Hüllermeier E (2003) Pairwise preference learning and ranking. In: Lavrac N, Gamberger D, Todorovski L, Blockeel H (eds) Proceedings of the European conference on machine learning (ECML 2003). Lecture notes in artificial intelligence, vol 2837. Springer, pp 145–156
10. Fürnkranz J, Hüllermeier E (eds) (2010) Preference learning. Springer, Berlin
11. Greco S, Matarazzo B, Słowiński R (1999) Rough approximation of a preference relation by dominance relations. Eur J Oper Res 117:63–83
12. Greco S, Matarazzo B, Słowiński R (2001) Rough sets theory for multicriteria decision analysis. Eur J Oper Res 129(1):1–47
13. Greco S, Matarazzo B, Słowiński R (2005) Decision rule approach. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis: state of the art surveys. Chap. 13. Springer, New York, pp 507–562

14. Greco S, Matarazzo B, Słowiński R (2005) Preference representation by means of conjoint measurement and decision rule model. In: Bouyssou D, Jacquet-Lagrèze E, Perny P, Słowiński R, Vanderpooten D, Vincke P (eds) Aiding decisions with multiple criteria—essays in honor of Bernard Roy. Kluwer, Boston, pp 263–313
15. Greco S, Matarazzo B, Słowiński R, Stefanowski J (2001) An algorithm for induction of decision rules consistent with the dominance principle. In: Ziarko W, Yao YY (eds) Rough sets and current trends in computing 2001. Lecture notes in artificial intelligence, vol 2005. Springer, Berlin, pp 304–313
16. Grzymała-Busse JW (1992) LERS—a system for learning from examples based on rough sets. In: Słowiński R (ed) Intelligent decision support. Handbook of Applications and Advances of the Rough Sets Theory. Kluwer, Dordrecht, pp 3–18
17. Grzymała-Busse JW (1997) A new version of the rule induction system LERS. Fundamenta Informaticae 31(1):27–39
18. Liu T-Y (2011) Learning to rank for information retrieval. Springer, Berlin
19. Roy B, Słowiński R (2013) Questions guiding the choice of a multicriteria decision aiding method. EURO J Decis Process 1(1):69–97
20. Saaty T (1980) The analytic hierarchy process. McGraw Hill, New York
21. Słowiński R, Greco S, Matarazzo B (2009) Rough sets in decision making. In: Meyers RA (ed) Encyclopedia of complexity and systems science. Springer, New York, pp 7753–7786
22. Słowiński R, Greco R, Matarazzo B (2014) Rough set based decision support. In: Burke EK, Kendall G (eds) Search methodologies: introductory tutorials in optimization and decision support techniques, Chap. 19, 2nd edn. Springer, New York, pp 557–609
23. Stefanowski J (2001) Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy. Rozprawy, vol 361. Wydawnictwo Politechniki Poznańskiej

On Things Not Seen

Marek Kimmel

Abstract Some statistical observations are frequently dismissed as “marginal” or even “oddities” but are far from such. On the contrary, they provide insights that lead to a better understanding of mechanisms which logically should exist but for which evidence is missing. We consider three case studies of probabilistic models in evolution, genetics and cancer. First, ascertainment bias in evolutionary genetics, arising when comparison between two or more species is based on genetic markers discovered in one of these species. Second, quasistationarity, i.e., probabilistic equilibria arising conditionally on non-absorption. Since evolution is also the history of extinctions (which are absorptions), this is a valid field of study. Third, inference concerning unobservable events in cancer, such as the appearance of the first malignant cell, or the first micrometastasis. The topic is vital for public health of aging societies. We try to adhere to mathematical rigor, but avoid professional jargon, with emphasis on the wider context.

1 Introduction

This essay attempts to persuade the Reader that statistical observations that may be dismissed as “marginal” or even “oddities” are far from such. On the contrary, they provide insights that lead to a better understanding of mechanisms which logically should exist but for which evidence is (and likely has to be) missing. To remain focused, we adhere to probabilistic models in evolution, genetics and cancer, disciplines in which the author claims expertise. The paper includes three case studies. First, ascertainment bias in evolutionary genetics, arising when comparison between two or more species is based on genetic markers discovered in one of these species.

M. Kimmel (✉)
Department of Statistics, Rice University, 6100 Main Street, Houston, TX 77005, USA
e-mail: kimmel@rice.edu

M. Kimmel
Systems Engineering Group, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_10

Second, quasistationarity, i.e., probabilistic equilibria arising conditionally on non-absorption. Since evolution is the history of extinctions (which are absorptions), this is a valid field of study. Third, inference concerning unobservable events in cancer, such as the appearance of the first malignant cell, or the first micrometastasis. The topic is vital for public health, particularly in aging societies. We try to adhere to mathematical rigor wherever needed and to provide references. Discussion concerns the wider context and philosophical implications.

2 Ascertainment Bias in Evolutionary Genetics

It has been observed that in evolutionary comparisons of Species 1 and 2, it is easy to err by using markers that were discovered in Species 1 and then sampled ("typed") in Species 1 and 2. Genetic markers have to exhibit among-individual variation to be useful and therefore if a marker is discovered in Species 1, then on the average it is more variable in Species 1 than in Species 2. Variability of markers serves as a proxy for the rate of nucleotide substitution, which in turn may be a proxy for the rate of evolution. For this reason, if Species 1 and 2 descend from a common ancestral species, such as Human and Chimpanzee, and markers discovered in Species 1 (Human, for example) are employed, then we may deduce that Human has been evolving faster than its sister species Chimpanzee, when in fact it has not [2, 7, 23]. One remedy for this effect (being a form of the ascertainment bias) is to also use markers discovered in Species 2 and compare the outcomes in both cases. However, how to analyze such data and what inferences might be drawn? Li and Kimmel [19] demonstrate that this is quite complicated and that conclusions may be far from obvious.

2.1 Microsatellite DNA and Divergence of Human and Chimpanzee

Microsatellite loci are stretches of repeated DNA motifs of length of 2–6 nucleotides. An example is a triplet repeat (motif of length 3) with allele length $X = 4$ (motif repeated 4 times)

$$\cdots|ACG|ACG|ACG|ACG|\cdots .$$

Mutations in such loci usually have the form of expansions or contractions occurring at a high rate, $\nu \sim 10^{-3}$–10^{-4} per generation. More specifically,

$$X \longrightarrow X + U \tag{1}$$

where U is an integer-valued random variable, at times constituting a Poisson process with intensity ν. Mutations in this Stepwise Mutation Model (SMM), mathematically form an unrestricted random walk (see e.g., [9]). Microsatellites are highly abundant in the genome. They are also highly polymorphic (variable). Applications of microsatellites include: forensics (identification), mapping (locating genes), and evolutionary studies.

A microsatellite locus can be considered to have a denumerable set of alleles indexed by integers. Two statistics can summarize the variability at a microsatellite locus in a sample of n chromosomes: The estimator of the genetic variance

$$\hat{V}/2 = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 / (n-1)\,, \tag{2}$$

where $X_i = X_i(t)$ is the length of the allele in the ith chromosome in the sample and $\overline{X}$ is the mean of the X_i

$$V(t) = E(\hat{V}) = E[(X_i - X_j)^2], \tag{3}$$

and X_i and X_j are exchangeable random variables representing the lengths of two alleles from the population [17]; and the estimator of homozygosity

$$\hat{P}_0 = (n\sum_{k=1}^{K} p_k^2 - 1)/(n-1), \tag{4}$$

where p_k denotes the relative frequency of allele k in the sample

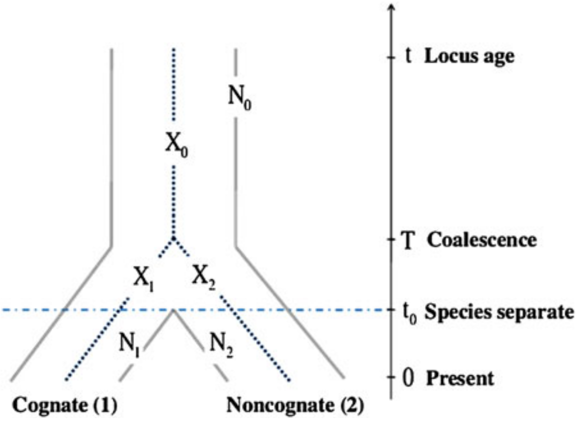


Fig. 1 Evolutionary history of a locus in two species. Demographic scenario employed in the mathematical model and simuPOP simulations. Notation: N_0, N_1, and N_2, effective sizes of the ancestral, cognate, and noncognate populations, respectively; X_0, X_1, and X_2, increments of allele sizes due to mutations in the ancestral allele, in chromosome 1 and in chromosome 2, respectively. **From** Ref. [19]

$$P_0(t) = E(\hat{P}_0) = \Pr[X_i(t) = X_j(t)]. \tag{5}$$

Random variables X_i are exchangeable but not independent.

Li and Kimmel [19] considered evolutionary history of a locus in two species. They employed the following demographic scenario in the mathematical model and simuPOP [20] simulations (Fig. 1). At time t before present (time is counted in reverse direction), a microsatellite locus is born in an ancestral species. At time t_0, the ancestral species splits into species 1 (called cognate) and species 2 (called non-cognate). Notation: N_0, N_1, and N_2, are effective lengths of the ancestral, cognate, and non-cognate populations, respectively; X_0, X_1, and X_2 are increments of allele lengths due to mutations in the ancestral allele, in chromosome 1 sampled at time 0 (present) from cognate population 1 and in chromosome 2, sampled from the non-cognate population 2.

2.2 Ascertainment Bias versus Drift and Mutation

In the random walk-like SMM model of mutation, a good measure of variability at a microsatellite locus is the length (repeat count) in a randomly sampled individual. Let us suppose that we discover a sequence of short motif repeats in the cognate species 1 and if its number of repeats Y_1 is greater or equal the threshold value x, we retain this microsatellite (we say we *discovered* it). Then we find a homologous microsatellite in species 2, i.e., microsatellite which is located in the same genomic region (technically, flanked by sequences of sufficient similarity), provided such microsatellite can be found. We take samples of microsatellite lengths from species 1 and 2, and consider their lengths to be realizations of random variables Y_1' and Y_2, respectively. We then consider the difference

$$D = E[Y_1'|Y_1 \geq x] - E[Y_2|Y_1 \geq x].$$

Other things being equal, D is a manifestation of the ascertainment bias and is likely to be positive. However, things may not be entirely equal. For example, if species 1 has a lower mutation rate than species 2, then its microsatellites will tend to have lower maximum length, which may reduce D. On the other hand, if, say, species 2 consistently has had a smaller population size, then genetic drift might have removed some of the variants and now species 2 microsatellites will have lower maximum length, which may inflate D. Li and Kimmel [19] carried out analytical and simulation studies of D under wide range of parameter values and obtained very good agreement of both techniques (Fig. 2). Briefly, as explained already, the observed difference D in allele lengths may be positive or negative depending on relative mutation rates and population sizes in the species 0 (ancestral), 1, and 2. In conclusion, mutation rate and demography may amplify or reverse the sampling (ascertainment) bias. Other effects were studied by different researchers. For example, Vowles and Amos [23] underscore the effects of upper bounds of repeat counts. An exhaustive discussion is found in Ref. [19].

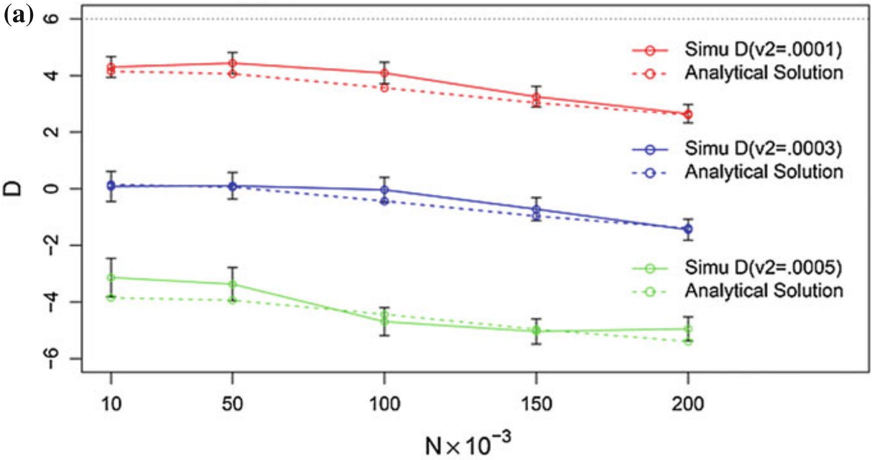


Fig. 2 Observed difference D in allele sizes may be positive or negative. Comparison of simuPOP simulations with computations based on Eq. (15). **a** Values of D for the basic parameter values $b_0 = b_1 = b_2 = b = 0.55$, $\nu_0 = \nu_1 = \nu = 0.0001$, $t_0 = 2 \times 10^5$ generations, and $t = 5 \times 10^5$ generations, with the effective sizes of all populations concurrently varying from 2×10^4 to 4×10^5 individuals and with mutation rates ν_2 varying from ν to 5ν. **b** Values of D for the basic parameter values $b_0 = b_1 = b_2 = b = 0.55$, $\nu_0 = \nu_2 = \nu = 0.0001$, $t_0 = 2 \times 10^5$ generations, and $t = 5 \times 10^5$ generations, with the effective sizes of all populations concurrently varying from 2×10^4 to 4×10^5 individuals and with mutation rates ν_1 varying from ν to 5ν (assuming 20 years per generation). **From** Ref. [19]

2.3 Hominid Slowdown and Microsatellite Statistics

Li and Kimmel [19] considered evidence for and against the so-called hominid slowdown (as discussed e.g., in Bronham et al. 1996), the observation that as the great apes become closer to the Human lineage, their nucleotide substitution rates (rates of point mutations in the genome) decrease. Consistent with this, Human and Human ancestors are expected to have slower substitution rates than Chimpanzee and its ancestors (following the divergence from the common ancestral species about 7 million years ago). Is this also true of microsatellite loci? Different molecular mechanisms shape these two types of mutations. Nucleotide substitutions result from random errors in DNA replication, which then may not be repaired, but also may lead to dysfunctional proteins which will be eliminated from the population by natural selection (as discussed e.g., in [10]). Microsatellite mutations, as explained already, result from replicase slippage. Most microsatellites are located in noncoding regions and therefore are considered selectively neutral.

The study [19] involves a reconstruction of the past demography of Human and its ancestors as well as hypothetical demography of Chimpanzee and its ancestors, including migrations of Human from its ancestral African territory and resulting population growth interrupted by recent glaciations and other events. Without getting into technical details, the conclusion is that microsatellite mutation rate is likely to be higher in Human than in Chimpanzee. It is interesting to observe that also the regulatory sites in the genome usually have the form of simple repeats (albeit interrupted) and vary quite considerably among species of mammals (as reviewed e.g., in Ref. [13]). It is possible to further hypothesize that evolution in higher mammals chose the path of regulation of gene expression as opposed to modification of the amino acid sequences in proteins; possible reason being that these latter might be too slow.

3 Quasistationarity in Genome Evolution

Let us consider an effect which is important if extinctions are indeed common in evolution. Suppose that a proliferating population has a random component of such nature that it leads any lineage to extinction with probability 1. On the other hand, proliferation is sufficiently fast to make up for extinction so that the non-extinct part may persist indefinitely. The long-term distribution of types of individuals in the population conditional on non-extinction, if such distribution exists, is called the quasistationary distribution. Quasistationarity in a more general sense has been studied by mathematicians for a long time; relevant literature has been collected by Pollet [21]. Here we will limit ourselves to an example from cell biology concerning gene amplification, based on an experiment pioneered by Schimke [22], with mathematical model developed by Kimmel and Axelrod [15] and then generalized by Kimmel [14] and Bansaye [3]. Let us notice that extinction causes information about evolution of the population to be scrambled. Therefore, if quasistationary distributions are interpreted as if they were ordinary stationary distributions, the conclusions may be paradoxical or misleading.

3.1 Gene Amplification in Cancer and Schimke's Experiments

One of the prevalent types of rearrangements in human cancer genome is gene amplification, i.e., increase of the number of gene copies in cells beyond the usual diploid complement. Some examples have been provided by [1], but the phenomenon is quite common, usually appearing under the guise of copy number variation (CNV; Fig. 3).

Classical experiments demonstrating gene amplification and its connection with drug resistance have been carried out in Schimke [22]. The gist of the experimental data can be described as follows. After passaging surviving cultured cells to ever increasing levels of metothrexate (MTX) over the period of the order of 10 Msec = 5 month, it was possible to evolve cells that were resistant to extremely high doses of MTX (Fig. 4). When the cells were put back into no MTX medium, they were observed to lose resistance within about 100 cell doublings (some cultures did not, but we sweep these under the rug for now).

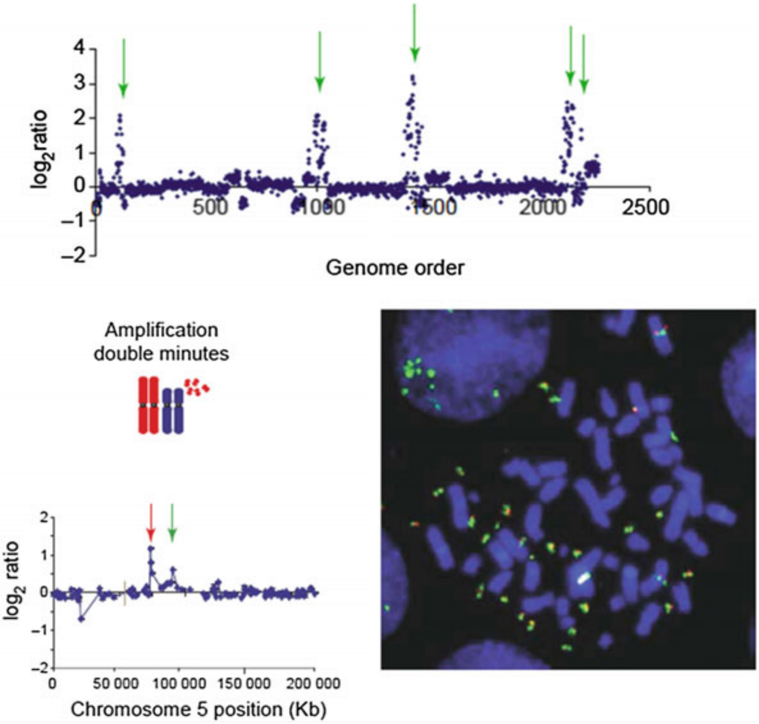
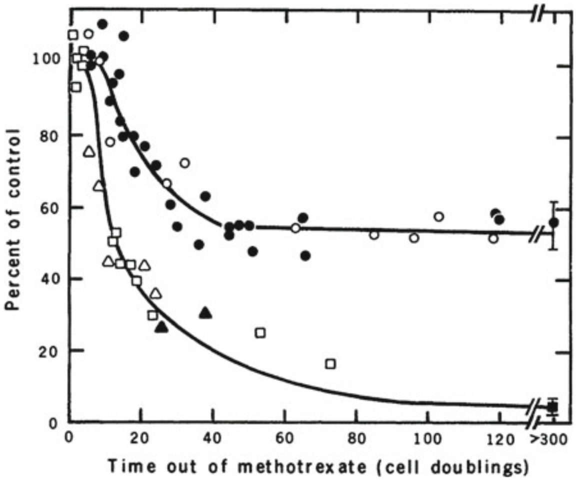


Fig. 3 Cytogenetics of gene amplification. Amplified DNA can be present in various forms including double minutes. A two-chromosome genome is depicted (*top* of the figure). Examples of array CGH copy number profiles (*bottom left*; plotted as the normalized log2 ratio) are shown with corresponding FISH pictures (*bottom right*) of the cells using BAC clones from the region of the amplicon indicated by the *red* and *green arrows*. Many red and green signals can be seen in the double minutes in a methotrexate-resistant human cell line. **From** Ref. [1]

Fig. 4 Loss of resistance in Schimke's experiments. Cells resistant to MTX are exposed to nonselective conditions. Some cell lines lose resistance completely (*circles*), while other only partially (*squares* and *triangles*). **From** Ref. [5]



Schimke discovered, using techniques available at that time, that the highly resistant cell had, besides the usual chromosomes, small extrachromosomal DNA elements roughly dicentric (he named them the “double minute chromosomes” or DM for short) that contained extra copies of the dihydrofolate reductase (DHFR) gene, that confers resistance to MTX [1]. It became clear that the increased resistance was due to amplification of the DHFR gene. But how did the amplified copies get there? Clearly a supercritical process of gene copy proliferation was at play. However, how did the cells know to multiply gene copies? The ghost of Lamarck knocked at the door.

3.2 Probabilists to Rescue

Fortunately for the common sense, Kimmel and Axelrod [15] conceived an idea consistent with the neo-Darwinian paradigm (despite appearances, this sentence is not necessarily an oxymoron). The hypothesis can be stated as follows:

- Increased resistance is correlated with increased numbers of gene copies on double minute chromosomes (DM).
- The number of DHFR genes on double minutes in a cell may increase or decrease at each cell division. This is because double minutes do not have centromeres, which are required to faithfully segregate chromosomes into progeny cells.
- The process of DM proliferation in cells is subcritical, since the DM do not efficiently replicate. Therefore cells grown in the absence of the drug gradually lose resistance to the drug, by losing extra gene copies.

The following model has been constructed by Kimmel and Axelrod [15].

- Galton-Watson process of gene amplification and deamplification in a randomly chosen line of descent (Fig. 5).
  - Double minute chromosomes replicate irregularly
  - Upon cell division, DMs are asymmetrically assigned to progeny cells.
- The process is subcritical, i.e., the average number of DMs at division is less than twice that number assigned to the cell at birth. This is consistent with imperfect replication and segregation of DMs.

Hypotheses of the model explain why, under nonselective conditions, the number of DMs per cell decreases which causes gradual loss of resistance (Fig. 5). In other words, zero is an absorbing state for the number of DMs. However, under selective conditions, only the cells with nonzero DM count survive. Therefore, conditionally on nonabsorption (non-extinction of the DMs), according to the Yaglom theorem for subcritical branching processes, the number of DMs per cell converges in distribution to a quasistationary distribution.

Specifically, suppose that proliferation of DMs from one cell generation to another, in a randomly selected ancestry line of cells is described by a Galton-Watson branch-

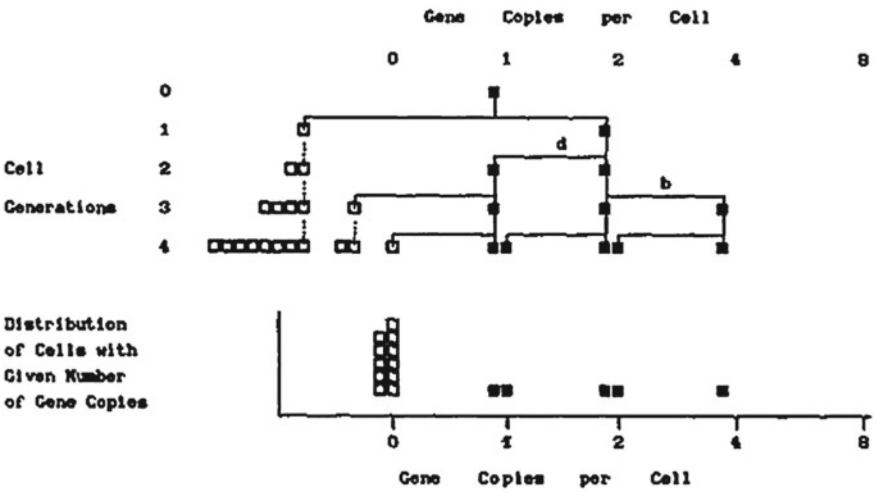


Fig. 5 A simplified view of gene amplification and deamplification process. Each cell with at least one gene copy can give rise to 2 progeny cells, each of which with probability b has amplified (doubled) count of DM gene copies, with probability d has deamplified (halved) count, or with probability $1 - b - d$, the same number. Halving of a single DM results in 0 DMs. Histogram at the bottom shows the resulting distribution of gene copies per cell in the fourth generation. **From** Ref. [15]

ing process with the number of "progeny" of a DM is a nonnegative integer random variable with generic probability generating function (pgf) $f(s)$, under the usual conditional independence hypotheses. As already noticed, this process is subcritical, i.e., $m = f'(1-) < 1$. Let Z_n denote the number of DMs in generation n and let $f_n(s)$ denote the pgf of Z_n.

Yaglom Theorem (see e.g., Theorem 4 in Kimmel and Axelrod [16]) *If* $m < 1$, *then* $\mathrm{P}[Z_n = j|Z_n > 0]$ *converges, as* $n \to \infty$ *to a probability function whose pgf* $\mathcal{B}(s)$ *satisfies the equation*

$$\mathcal{B}[f(s)] = m\mathcal{B}(s) + (1 - m).$$

Also,

$$1 - f_n(0) \sim \frac{m^n}{\mathcal{B}'(1-)},\ n \to \infty.$$

Yaglom limit is also an example of a *quasistationary* distribution, say $\mu(x)$, which in a general Markov chain can be defined via the following condition

$$\mu(x) = \frac{\sum_{y\geq 1} \mu(y)\mathrm{P}_y[X(t) = x]}{\sum_{y\geq 1} \mu(y)\mathrm{P}_y[X(t) \neq 0]},$$

where $\mathrm{P}_y[X(t) = x]$ is the transition probability matrix.

Let us suppose now that cell population has been transferred to MTX-free medium at generation $n = N$. Based on the Yaglom Theorem, the fraction of resistant cells decreases roughly geometrically

$$1 - f_n(0) \sim \frac{m^{n-N}}{\mathcal{B}'(1-)}, \; n > N,$$

while $\{Z_n | Z_n > 0\}$ remains unchanged. Moreover, if $2m > 1$, then the net growth of the resistant population is observed also at the selection phase ($n \leq N$).

Loss of DMs in non-selective conditions has been visualized experimentally [5]. Population distribution of numbers of copies per cell can be estimated by flow cytometry. Proportion of cells with amplified genes decreases with time (Fig. 6). Shape of the distribution of gene copy number in the subpopulation of cells with amplified genes appears unchanged as resistance is gradually lost.

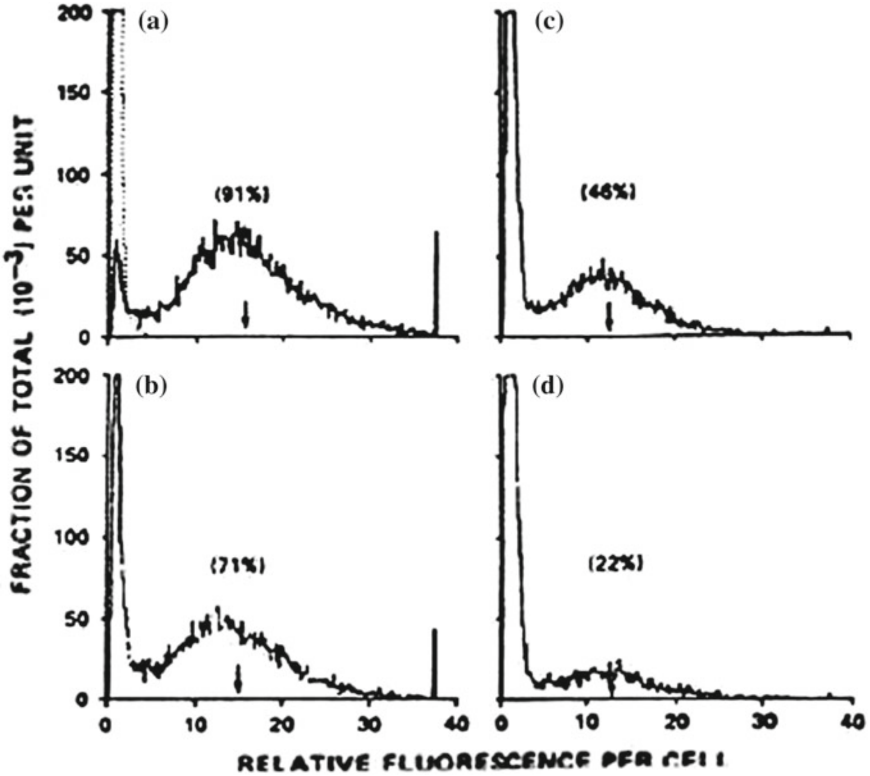


Fig. 6 Loss of resistance visualized by flow cytometry. Population distribution of numbers of copies per cell can be estimated by flow cytometry. Proportion of cells with amplified genes decreases with time. Shape of the distribution of gene copy number in the subpopulation of cells with amplified genes appears unchanged as resistance is gradually lost. **From** Ref. [5]

An important finding is that if $2m > 1$, i.e., if absorption is not too fast, then cell proliferation outweighs the loss of cell caused by the selective agent (MTX) and the resistant subpopulation grows in absolute numbers also under selective conditions (when $n \leq N$; details in the original paper and the book).

A more general mathematical model of replication of "small particles" within "large particles" and of their asymmetric division ("Branching within branching") has been developed by Kimmel [14] and followed up by Bansaye [3]. It is interesting to notice that quasistationary distributions are likely to generate much heterogeneity. An example is provided by large fluctuations of the critical Galton-Watson process before extinction; see Wu and Kimmel [24].

3.3 Quasistationarity and Molecular Evolution

An observation can be made that trends observed in molecular evolution can be misleading, if they are taken at their face value and without an attempt to understand their underlying "mechanistic" structure. It may be concluded, looking at the evolution of resistance in cells exposed to MTX that there exists something in the MTX that literally leads to an increase of the number of DM copies. So, gene amplification is "induced" by MTX. Only after it is logically deduced that DMs have to undergo replication and segregation and assuming that both these processes are less orderly in DMs than in the "normal" large chromosomes, the conclusion concerning the true nature of the process (selection superimposed on *subcritical* branching) follows by the laws of population genetics.

4 Unobservables in Cancer

Early detection of cancer by mass screening of at risk individuals remains one of the most contentious issues in public health. We will mainly use lung cancer (LC) as an example. The idea is to identify the "at risk" population (smokers in the LC case), and then to apply an "early detection" procedure (CT-scan in the LC case), periodically, among the members of the "at risk" population. By treating the early (and implicitly, curable) cases discovered this way, a much higher cure rate is assured than that of spontaneously appearing symptomatic cases. Is this reasoning correct? Two types of arguments have been used to question the philosophy just described. On one hand, part of the early detection may constitute overdiagnosis. Briefly, by the effect known from the renewal theory, a detection device with less than perfect sensitivity, placed at a fixed point in time and confronted with examined cases "flowing in time", preferentially detects cases of longer duration, i.e. those for which the asymptomatic early disease is more protracted. This effect is known as the length-biased sampling (discussion in Ref. [12]). Its extreme form, called overdiagnosis, causes detection of cases that are so slow that they might show only at autopsy, or cases which look like cancer but do not progress at all. Overdiagnosis, if it were frequent, would invalidate early detection: a large number of “early non-cancers” would be found and unnecessarily treated, causing increased morbidity and perisurgical mortality, without much or any reduction in LC death count.

On the other hand, the following scenario is possible, which also may invalidate screening for early detection, although for an opposite reason. If it happens that LC produces micrometastases, which are present when the primary tumor is of submillimeter size, then detection of even 2–3 mm tumors (achievable using CT) is futile, since the micrometastases progress and kill the patient whether the primary tumor has been detected or not.

How to determine if screening reduces number of LC deaths? The orthodox biostatistics approach is “empirical”. It consists of designing a two-arm RCT (screened versus non-screened high risk individuals) and comparing numbers of LC deaths in the two arms. This methodology is statistically sound, but it may be considered unethical. Patients in the control arm are denied potentially life-saving procedures. Those in the screened arm do not necessarily take advantage of the current state-of-art technology. Two sources of reduced contrast are: noncompliance in the screened arm and/or “voluntary” screening in the control arm. It has been claimed that the results of the Mayo Lung Project (MLP) 1967–1981 trial, which influenced recommendations not to screen for LC by chest X ray were simply due to lack of power to demonstrate mortality reduction by 5–10% which might be achievable using X ray screening [12]. Finally, the National Lung Screening Trial (NLST) in the USA, in which around 50,000 smokers took part, demonstrated that a series of three annual screenings followed by treatment of detected cases reduced mortality by about 20 %. It has to be noted, that predictions of similar magnitude reduction obtained using modeling [18] have been almost universally disregarded by the medical community.

The NLST has left as many questions unanswered as it did answer. One of them is the choice of the “best” high-risk group for LC screening. Given limited resources, how to allocate them to subgroups of screenees so that the efficacy of a mass screening program is maximized. Even if the meaning of the term “efficacy” is clarified, it is still unknown who should be screened. Are these the heaviest smokers, the smokers who have smoked for the longest period of time, individuals with family history of lung cancer, or those with impaired DNA-repair capacity [11]? At what age does it make sense to start screening and how often should the individuals be screened? Answers to these questions require knowledge of the natural course of disease, which is exactly what is not observable (Fig. 7).

Arguably, modeling can help. If a model of carcinogenesis, tumor growth and progression (i.e., nodal and distant metastases) is constructed and validated and models of early detection and post-detection follow-up are added to it, then various scenarios of screening can be tested in silico. Another use of modeling is less utilitarian, but equally important. It can be called the inverse problem: How much is it possible to say about the course of cancer based on snapshots including the disease characteristics at detection? To what extent is the size of the primary tumor predictive of the progression of the disease? In [6] some inferences of this type have been made (Fig. 8).

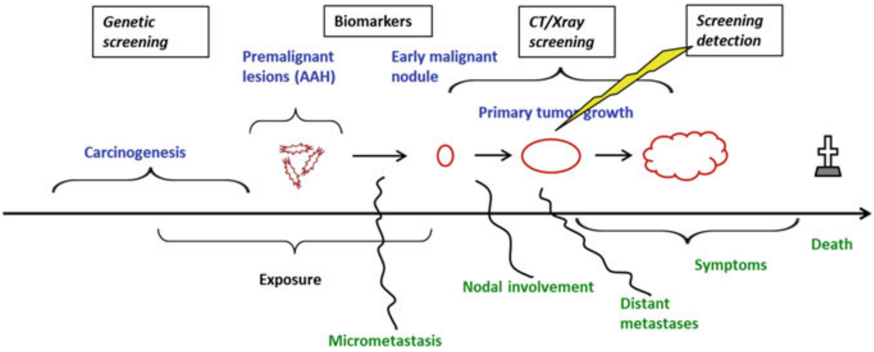
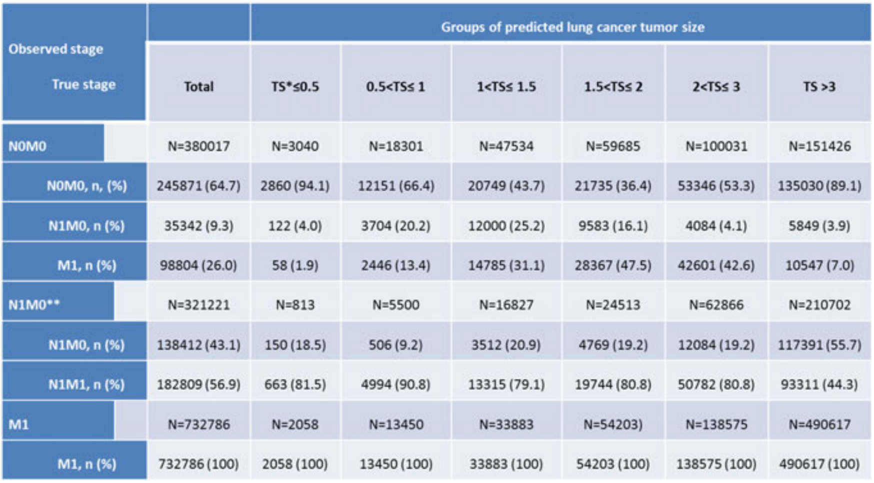


Fig. 7 Time lines of cancer progression and detection



Observed stage True stage	Total	Groups of predicted lung cancer tumor size					
		TS*≤0.5	0.5<TS≤ 1	1<TS≤ 1.5	1.5<TS≤ 2	2<TS≤ 3	TS >3
N0M0	N=380017	N=3040	N=18301	N=47534	N=59685	N=100031	N=151426
N0M0, n, (%)	245871 (64.7)	2860 (94.1)	12151 (66.4)	20749 (43.7)	21735 (36.4)	53346 (53.3)	135030 (89.1)
N1M0, n (%)	35342 (9.3)	122 (4.0)	3704 (20.2)	12000 (25.2)	9583 (16.1)	4084 (4.1)	5849 (3.9)
M1, n (%)	98804 (26.0)	58 (1.9)	2446 (13.4)	14785 (31.1)	28367 (47.5)	42601 (42.6)	10547 (7.0)
N1M0**	N=321221	N=813	N=5500	N=16827	N=24513	N=62866	N=210702
N1M0, n (%)	138412 (43.1)	150 (18.5)	506 (9.2)	3512 (20.9)	4769 (19.2)	12084 (19.2)	117391 (55.7)
N1M1, n (%)	182809 (56.9)	663 (81.5)	4994 (90.8)	13315 (79.1)	19744 (80.8)	50782 (80.8)	93311 (44.3)
M1	N=732786	N=2058	N=13450	N=33883	N=54203)	N=138575	N=490617
M1, n (%)	732786 (100)	2058 (100)	13450 (100)	33883 (100)	54203 (100)	138575 (100)	490617 (100)

Fig. 8 Distributions of occult nodal and distant metastases in the simulated lung cancer patients (1988–1999) with stage N0M0, N1M0 and M1 stratified by tumor size. *TS, Primary tumor size (cm) in diameter **In SEER data, 7208 were N0M1, which is 9.7 % of 74109 that had N and M staged. This stage is not modeled. **From** Ref. [6]

Figure 8 depicts distributions of undetected (so-called occult) nodal and distant metastases in the simulated lung cancer patients, fitting demographic and smoking patterns of the SEER database 1988–1999, detected with stage N0M0, N1M0 and M1, stratified by primary tumor size. N0 and N1 correspond to the absence and presence of lymph node metastasis, and M0 and M1 to the absence and presence of distant metastasis, respectively. In other words, modeling allows to estimate how many of lung cancers detected as belonging to a given category, in reality belong to different, prognostically less favorable, categories. The results show some unexpected trends. The most important are the three top rows of Fig. 8, which concern tumors detected without nodal or distant metastases (N0M0). These tumors, on the face of things, offer best prognosis. Model predictions confirm this intuition, up to a point. Indeed up to the primary tumor size of about 1 cm, more than 50 % of apparent N0M0 tumors are indeed N0M0. If they are detected at larger sizes, then only a minority are truly N0M0, and the rest have occult metastases. So, if these tumors below 1 cm are removed, there is a good chance the patient is cured. But, surprisingly, there is another turning point. At sizes above 2.5–3 cm, again majority of tumors are N0M0. Similar, though not as distinctive trend is present when we consider tumors detected as N1M0. Therefore, if a very large tumor is discovered without apparent nodal and distant metastasis and it is resectable, then the suggestion is that it might be resected for cure.

The explanation for this peculiar pattern is that if the rates of growth and progression (metastasizing) of tumors are distributed, then detection is "cutting out windows" in the distributions, through which the tumor population is observed. In the large primary tumor size category with no metastases observed, we deal with the fast growing, slowly metastasizing subset. Other subpopulations simply present with metastasis when the primary tumor is large, become symptomatic and quickly progress to death. So, active detection leads to biased TNM distributions, with the bias sometimes being non-intuitive.

Mathematical models of the NLST trial predicted its outcome in two publications, one in 2004 [18] and the other in 2011 ([8]; submitted for publication before the NLST outcome was announced), using two different modeling approaches. As stated already, at that time these papers were universally ignored.

5 Discussion

What is the role and use of statistics as a profession (science?) and of statisticians as professionals (scientists?). In minds of other scientists (physicists, biologists or physicians) statistics is mainly perhaps a useful, but strictly confirmatory field. What is expected of a collaborating statistician is the "p-value" or the sample size needed to obtain a given "power" of a test as required by the funding agencies. However, one may reflect on how many useful and deep scientific concepts and techniques are statistical in nature. Some of them have been revolutionary. We may list some with biological applications: Fluctuation Analysis (FA) in cell biology, Moolgavkar-Knudson (M-K) model of carcinogenesis, Wright-Fisher (W-F) model in population genetics, Capture-Recapture (C-R) method in ecology, Maximum Likelihood (ML) and Least Squares (LS) methods in molecular phylogenetics, and other. However, let us notice that these methods are based on models that include structural features of the biological nature of the phenomenon in question. Some of these are unobservable, such as mutations in cells in FA, stage transition times in M-K, segregation of chromosomes to progeny in W-F, collections of individuals in C-R and ancestral nodes in phylogenetics.

Arguably, statistics is most useful, when it considers phenomena in a "gray zone" such as inference on the unseens, i.e., processes that we believe are real, but which cannot be directly observed. Three phases of scientific inquiry, are usually present:

1. Initially, when there is little or no data; the unseens are not suspected to exist,
2. Existence of the unseens is revealed through progress in data collection and logical analysis,
3. Further progress may lead to resolution of the unseen by a reductionist approach.

Examples considered in the essay involve analyses in Phase 2. Each involves unseens that may become observable at some time. Also, each required construction of a new model based on inferred biology of the process. In addition, each of the examples includes a statistical sampling mechanism, which introduces a bias (we may call it the ascertainment bias). The role of the model is among other, to understand and counter the bias. Arguably, this is the true purpose of statistical analysis.

References

1. Albertson DG (2006) Gene amplification in cancer. Trends Genet 22:447–455
2. Amos W et al (2003) Directional evolution of size coupled with ascertainment bias for variation in drosophila microsatellites. Mol Biol Evol 20:660–662
3. Bercu B, Blandin V (2014) Limit theorems for bifurcating integer-valued autoregressive processes. Statistical inference for stochastic processes, pp 1–35
4. Bromham L, Rambaut A, Harvey PH (1996) Determinants of rate variation in mammalian DNA sequence evolution. J Mol Evol 43:610–621
5. Brown PC, Beverley SM, Schimke RT (1981) Relationship of amplified dihydrofolate reductase genes to double minute chromosomes in unstably resistant mouse fibroblast cell lines. Mol Cell Biol 1:1077–1083
6. Chen X et al (2014) Modeling the natural history and detection of lung cancer based on smoking behavior. PloS one 9(4):e93430
7. Cooper G, Rubinsztein DC, Amos W (1998) Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. Hum Mol Genet 7:1425–1429
8. Foy M et al (2011) Modeling the mortality reduction due to computed tomography screening for lung cancer. Cancer 117(12):2703–2708
9. Goldstein DB, Schlotterer C (1999) Microsatellites: evolution and applications, pp 1–368
10. Gorlov IP, Kimmel M, Amos CI (2006) Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. Hum Mol Genet 15:1143–1150
11. Gorlova OY et al (2003) Genetic susceptibility for lung cancer: interactions with gender and smoking history and impact on early detection policies. Hum Hered 56:139–145
12. Gorlova OY, Kimmel M, Henschke C (2001) Modeling of long-term screening for lung carcinoma. Cancer 92:1531–1540
13. Iwanaszko M, Brasier AR, Kimmel M (2012) The dependence of expression of NF-κB-dependent genes: statistics and evolutionary conservation of control sequences in the promoter and in the 3' UTR. BMC Genomics 13:182
14. Kimmel M (1997) Quasistationarity in a branching model of division-within-division. In: Classical and modern branching processes (Minneapolis, MN, 1994), pp 157–164. Springer, New York. IMA Vol Math Appl 84

15. Kimmel M, Axelrod DE (1990) Mathematical models of gene amplification with applications to cellular drug resistance and tumorigenicity. Genetics 125:633–644
16. Kimmel M, Axelrod DE (2015) Branching processes in biology (2nd edn, extended). Springer, Heidelberg
17. Kimmel M et al (1996) Dynamics of repeat polymorphisms under a forward-backward mutation model: within-and between-population variability at microsatellite loci. Genetics 143:549–555
18. Kimmel M, Gorlova OY, Henschke CI (2004) Modeling lung cancer screening. Recent advances in quantitative methods in cancer and human health risk assessment, pp 161–175
19. Li B, Kimmel M (2013) Factors influencing ascertainment bias of microsatellite allele sizes: impact on estimates of mutation rates. Genetics 195:563–572
20. Peng B, Kimmel M (2005) simuPOP: a forward-time population genetics simulation environment. Bioinformatics 21:3686–3687
21. Pollett PK (2014) Quasi-stationary distributions: a bibliography. http://www.maths.uq.edu.au/~pkp/papers/qsds/qsds.pdf
22. Schimke RT (ed) (1982) Gene amplification, vol 317. Cold Spring Harbor Laboratory, New York
23. Vowles EJ, Amos W (2006) Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. Mol Biol Evol 23:598–607
24. Wu X, Kimmel M (2010) A note on the path to extinction of critical Markov branching processes. Statist Probab Lett 80:263–269

Network Capacity Bound for Personalized Bipartite PageRank

Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, Robert A. Kłopotek and Elżbieta A. Kłopotek

Abstract In this paper a novel notion of Bipartite PageRank is introduced and limits of authority flow in bipartite graphs are investigated. As a starting point we simplify the proof of a theorem on personalized random walk in unimodal graphs that is fundamental to graph nodes clustering. As a consequence we generalize this theorem to bipartite graphs.

Keywords Bipartite graphs · Social networks · PageRank

1 Introduction

The notion of the PageRank as a measure of importance of a web page was introduced in [11]. Since then a large flow of publications on this topic emerged, starting with methods of computation [2] and numerous applications (Web page ranking, client and seller ranking, clustering, classification of web pages, word sense disambiguation, spam detection, detection of dead pages etc.) and variations (personalized PageRank, topical PageRank, Ranking with Back-step, Query-Dependent PageRank etc.), [7].

This is an extended version of the paper presented at the AI2012 Conference, Sept. 20th, 2012, Warsaw, Poland.

M.A. Kłopotek (✉) · S.T. Wierzchoń
Institute of Computer Science of Polish Academy of Sciences, Warszawa, Poland
e-mail: klopotek@ipipan.waw.pl

R.A. Kłopotek
International PhD. Programme at ICS PAS, Warszawa, Poland

M.A. Kłopotek
Institute of Computer Science of Natural and Human Sciences University, Siedlce, Poland

E.A. Kłopotek
m-Bank, Warszawa, Poland

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_11

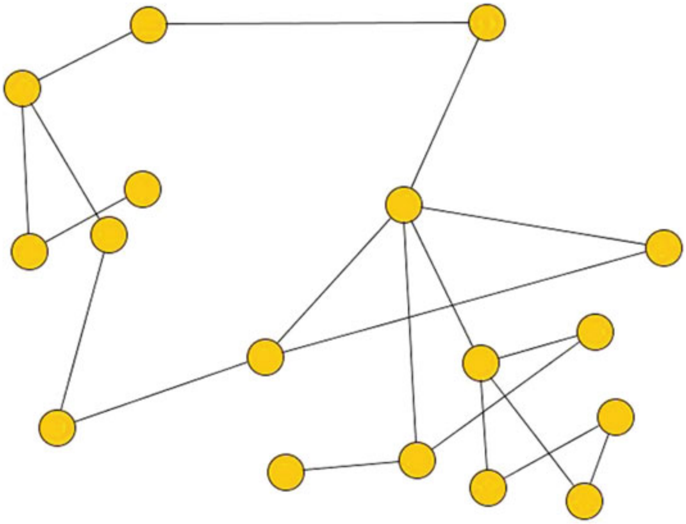


Fig. 1 An unoriented network

In this paper our attention is focused on a certain aspect of personalized PageRank, related to its usage as a way to cluster nodes of an undirected graph[1] (see Fig. 1). While a number of algorithms has been developed based on PageRank variations with the concept that a group of pages forms a cluster when it is unlikely to be left by a random walker, the fundamental theorem on this issue seems to have quite a complex proof—see e.g. [3, Lemma 2].

In this paper we will make an attempt to simplify it (Sect. 2) and further we will generalize it to bipartite graphs (Sect. 3).

Our interest in bipartite graphs arouse due to the fact that such graphs are of particular interests in a number of application domains where the relationships between objects may be conveniently represented in the form of a graph. The success story of PageRank prompts the researchers to apply it also to those graphs e.g. describing mutual evaluations of students and lecturers [9], reviewers and movies in a movie recommender systems, or authors and papers in scientific literature or queries and URLs in query logs [4], or performing image tagging [1]. However, these attempts seem to be somehow counterproductive because the PageRank was explicitly designed to remove periodicity from the graph structure, whereas the bipartite graphs have explicitly this kind of structure. Therefore a suitable generalization of PageRank to such structures is needed in order to retain both advantages of the bipartite graph representation and those of PageRank. This paper can be seen as a contribution in

[1] An unoriented graph may serve as the representation of relationships spanned by a network of friends, telecommunication infrastructure or street network of a city, etc.

this direction. Please note that Theorem 4 in [10] states that "PageRank's algorithm will not work on bipartire graphs". We provide a method allowing to overcome this inconvenience.

2 Unimodal PageRank

Let us restrict to one of the many interpretations of PageRank as the probability ("in the limit") that a knowledgeable but mindless random walker (see Fig. 2) will encounter a given Web page. Knowledgeable because he knows the addresses of all the web pages. Mindless because he chooses a next page to visit without paying attention to any hints on its contents. So upon entering a particular web page, if it has no outgoing links, the walker jumps to any Web page with uniform probability. If there are outgoing links, he chooses with uniform probability one of the outgoing links and goes to the selected web page, unless he gets bored. If he gets bored (which may happen with a fixed probability ζ on any page), he jumps to any web page with uniform probability. Thus ζ is referred to as *teleportation probability* or *dumping factor*. This variant of PageRank shall be called *traditional uniform PageRank*. Careful analysis of its properties with particular emphasis on the case $\zeta \to 1$ can be found in [5].

At the other extreme we can consider a mindless page-u-fan random walker who is doing exactly the same, but in case of a jump out of boredom he does not jump to any page, but to the page u. A page ranking obtained in this way is called *single-page*

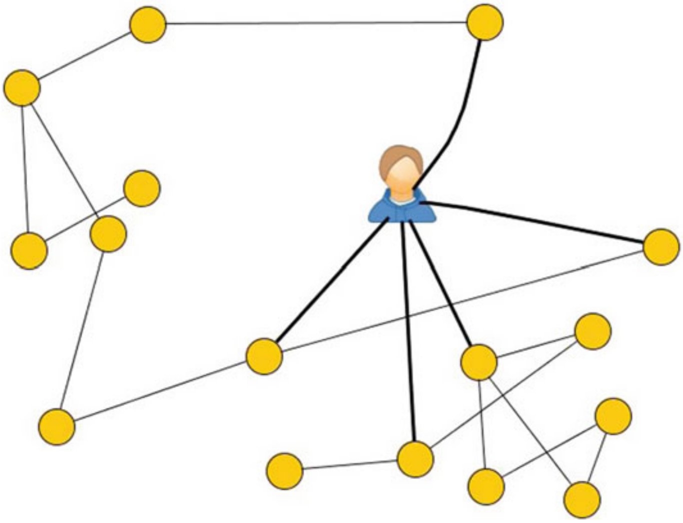


Fig. 2 Random walker interpretation of PageRank

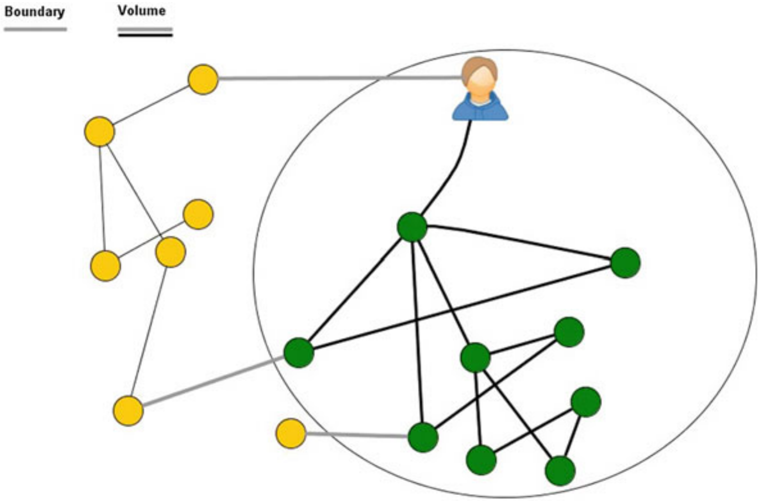


Fig. 3 A preferred group of pages of a random walker

personalized PageRank. Its applications to classification of the users of a social network site into groups are discussed e.g. in [6].

There is an interesting relationship between the knowledgeable and the page-fan walkers: if there exists one page-fan for each web page then the traditional uniform PageRank vector of the knowledgeable walker is the average of single-page personalized PageRank vectors of all these one page-fan walkers (Fig. 3).

Also there are plenty possibilities of other mindless walkers between these two extremes. For example once the walker is bored, he can jump to a page from a set U with a uniform probability or with probability proportional to the out-degree of the pages from this set. A detailed treatment of these topics can be found e.g. in [8] to which unacquainted reader is warmly referred.

In any of the cases the PageRank is interpreted as the probability "in the limit" of a random walker reaching a web page. This probability can be considered as a kind of "authority" of the page. We can assign also to the edges the probability that a walker will pass this edge. The probabilities assigned to edges can be viewed as a "flow of authority".

So let us introduce some notation. By $\mathbf{r}$ we will denote a (column) vector of ranks: r_j will mean the PageRank of page j. All elements of $\mathbf{r}$ are non-negative and their sum equals 1.

Let $\mathbf{P} = [p_{ij}]$ be a matrix such that if there is a link from page j to page i, then $p_{i,j} = \frac{1}{outdeg(j)}$, where $outdeg(j)$ is the out-degree of node j.[2] In other words, $\mathbf{P}$ is column-stochastic matrix satisfying $\sum_i p_{ij} = 1$ for each column j. If a node had an

[2] For some versions of PageRank, like TrustRank $p_{i,j}$ would differ from $\frac{1}{outdeg(j)}$ giving preferences to some outgoing links over others. We are not interested in such considerations here.

out-degree equal 0, then prior to construction of **P** the node is replaced by one with edges outgoing to all other nodes of the network.

Under these circumstances we have

$$\mathbf{r} = (1 - \zeta) \cdot \mathbf{P} \cdot \mathbf{r} + \zeta \cdot \mathbf{s} \tag{1}$$

where **s** is the so-called "initial" probability distribution (i.e. a column vector with non-negative elements summing up to 1) that is also interpreted as a vector of Web page preferences.

For a knowledgeable walker (case of traditional uniform PageRank) for each node j of the network $s_j = \frac{1}{|N|}$, where $|N|$ is the cardinality of the set of nodes N constituting the network. For a page-u-fan (single-page personalized PageRank) we have $s_u = 1$, and $s_j = 0$ for any other page $j \neq u$.

Let us introduce now *multipage uniform personalized PageRank*. It corresponds to a random walker that once he is bored, he jumps to any of the pages from a set U with uniform probability—one may say the walker is a uniform-set-U-fan. we get

$$s_j = \begin{cases} \dfrac{1}{|U|} & \text{if } j \in U \\ 0 & \text{otherwise} \end{cases}, \quad j = 1, \ldots |N|$$

In this paper we will be interested in *multipage hub-oriented personalized PageRank*. It corresponds to a random walker that, once bored, jumps to any of the pages from a set U with probability proportional to its out-degree—one may say the walker is a hub-oriented-set-U-fan.

$$s_j = \begin{cases} \dfrac{outdeg(j)}{\sum_{k \in U} outdeg(k)} & \text{if } j \in U \\ 0 & \text{otherwise} \end{cases}, \quad j = 1, \ldots |N| \tag{2}$$

Instead of looking at a random walker or at authority flow "in the limit", we can look at the Web as a pipe-net through which the authority is flowing in discrete time steps, starting from the initial distribution defined by some $\mathbf{r}^{(0)}$ vector.

In single time step a fraction ζ of the authority of a node j flows into so-called *super-node*, and the fraction $\frac{1-\zeta}{outdeg(j)}$ is sent from this node to each of its children in the graph. After the super-node has received authorities from all the nodes, it redistributes the authority to all the nodes in fractions defined in the vector **s**. We assume that the authority circulates lossless (we have a kind of a closed loop here).

Beside this, as was proven in many papers, we have to do here with a self-stabilizing process (under some conditions). Starting with any stochastic vector $\mathbf{r}^{(0)}$ and applying the operation

$$\mathbf{r}^{(n+1)} = (1 - \zeta) \cdot \mathbf{P} \cdot \mathbf{r}^{(n)} + \zeta \cdot s$$

the series $\{\mathbf{r}^{(n)}\}$ will converge to $\mathbf{r}$, the dominating eigenvector[3] of the matrix $\mathbf{P}$ from Eq. (1) (i.e. $\mathbf{r}$ is the eigenvector corresponding to the maximal eigenvalue $\lambda = 1$ of the matrix $\mathbf{P}$).

Subsequently let us consider only connected graphs (one-component graphs) with symmetric links, i.e. undirected graphs. An unoriented edge between nodes means that there are oriented edges in both directions. Hence for each node j the relationships between in- and out-degrees are:

$$indeg(j) = outdeg(j) = deg(j)$$

Let us pose the question: how is the PageRank of U-set pages related to the PageRank of other pages (that is those pages where there are no jumps out of being bored)?

First, let us note some obvious cases.

(a) Assume that $\zeta = 0$, that is the random walker does not get bored. Let $K = \sum_{j \in N} deg(j)$ denote the number of outgoing edges of the whole undirected network (K is twice as large as the number of unoriented edges of the network). Then for each j: $r_j = \frac{deg(j)}{K}$, which is proved easily when one looks at the network as a network of channels through which the authority flows from one node to the other. Because the super-node does not play any role, and with the above-mentioned $\mathbf{r}$ each channel is filled with the same amount of authority $\frac{1}{K}$, that is in each bi-directional channel the flow is compensated. Hence such a PageRank vector remains stable—it is the fixed point of Eq. (1). Note that self-stabilizing is not guaranteed here.

(b) On the other extreme, if $\zeta = 1$ (only jumps through super-node) then $\mathbf{r} = \mathbf{s}$ (stability guaranteed).

Let us note that for $\zeta > 0$ the authority accumulated by the super-node equals exactly ζ. Particularly, this amount of the authority is redistributed to all the "fan" pages $U \subset N$.

Let us discuss now a hub-oriented-set-U-fan defined in Eq. (2). Assume for a moment that the set U is identical with N, and that at a time-point t the authority distribution is

$$r_j^{(t)} = \frac{deg(j)}{K} \tag{3}$$

Consider the next moment $t + 1$. Authority flew from the node j to the super-node amounting to $\zeta \frac{deg(j)}{K}$, and into each (outgoing) link $(1 - \zeta)\frac{1}{K}$. The very same node j gets from the super-node authority amounting to $\zeta \frac{deg(j)}{K}$, and from each (ingoing) link $(1 - \zeta)\frac{1}{K}$. Hence $\mathbf{r}^{(t+1)} = \mathbf{r}^{(t)}$ so the $\mathbf{r}^{(t)}$ defined in Eq. (3) is our solution of the Eq. (1).

For nonzero ζ the stabilization is guaranteed.

Let us now turn to the situation where U is only a proper subset of N, and let us ask: "How much authority from outside of U can flow into U via super-node at the

[3] Called also *principal eigenvector*.

point of stability?" Let us denote by p_o the total mass of authority contained in all the nodes outside of U. Then our question concerns the quantity $p_o\zeta$. Note that this is the same amount of authority that leaves U via links outgoing from this set to its complement (as a net balance). It is claimed that, see e.g. [3]:

Theorem 1 *For the multipage hub-oriented personalized PageRank we have*

$$p_o\zeta \leq (1-\zeta)\frac{|\partial(U)|}{Vol(U)}$$

where $\partial(U)$ is the set of edges leading from U to the nodes outside of U (the so-called "edge boundary of U"), hence $|\partial(U)|$ is the cardinality of the boundary, and $Vol(U)$, called volume or capacity of U, is the sum of out-degrees of all nodes from U.

Before starting the proof let us stress the importance of this theorem for the task of clustering. Some clustering algorithms assume that the set U is a good cluster if the amount of authority $p_o\zeta$ leaving it via outgoing links is low. But this theorem states that this is not the full story. If you take a bigger subgraph then a chance of authority leaving it will be reduced no matter how the internal structure looks like. So in fact the cluster value should be considered with respect to this natural limitation—"the worst case" of internal link structure and not just the absolute value.

Proof The idea of the proof is as follows:

1. We want to find a vector $\mathbf{r}^b$ of authorities such that if $\mathbf{r}^{(t)} \leq \mathbf{r}^b$ at a time point t, then this inequality is true for any $t' > t$.
2. Then we shall show that such a vector $\mathbf{r}^{(t)}$ of authorities really exists, for which $\mathbf{r}^{(t)} \leq \mathbf{r}^b$ holds.
3. Next we shall demonstrate that starting with any initial conditions the vector $\mathbf{r}^{(t)} \leq \mathbf{r}^b$ is reachable.
4. The proof finishes when we show that the limiting vector $\mathbf{r}^b$ allows us to conclude that the claim of the theorem is valid.

ad.1. To find the distribution $\mathbf{r}^{(t')}$ for $t' > t$ we state that if in none of the links the passing amount of authority will exceed

$$\gamma = (1-\zeta)\frac{1}{\sum_{k\in U} deg(k)}$$

then at any later time point $t' > t$ the inequality $r_j^{(t')} \leq r_j^{(t)}$ holds at any node $j \in U$.

To justify this statement note that if a node $j \notin U$ gets via links $l_{j,1}, \ldots, l_{j,deg(j)}$ the authority amounting to

$$a_{l_{j,1}} \leq \gamma, \ldots, a_{l_{j,deg(j)}} \leq \gamma$$

then it accumulates

$$\mathfrak{a}_j = \sum_{k=1}^{deg(j)} a_{j,k} \leq \gamma \cdot deg(j)$$

of total authority, and in the next time step the following amount of authority flows out through each of these links:

$$(1-\zeta)\frac{\mathfrak{a}_j}{deg(j)} \leq \gamma(1-\zeta) \leq \gamma$$

If a node $j \in U$ gets via incoming links $l_{j,1}, \ldots, l_{j,deg(j)}$ the authority amounting to $a_{l_{j,1}} \leq \gamma, \ldots, a_{l_{j,deg(j)}} \leq \gamma$ then, due to the authority obtained from the super-node equal to $\mathfrak{b}_j = \zeta \frac{deg(j)}{\sum_{k \in U} deg(k)} = deg(j)\gamma\frac{\zeta}{1-\zeta}$, in the next step through each link the authority amounting to

$$(1-\zeta)\frac{\mathfrak{a}_j}{deg(j)} + (1-\zeta)\frac{\mathfrak{b}_j}{deg(j)} \leq \gamma(1-\zeta) + \gamma\frac{\zeta}{1-\zeta}(1-\zeta)$$
$$= \gamma(1-\zeta) + \gamma\zeta = \gamma$$

flows out.

So if already at time point t the authority flowing out through any link from any node did not exceed γ (they were equal to it or to zero), then this property will hold (by induction) forever. This phenomenon is illustrated on Fig. 4. Here U is a subset consisting of 10 nodes randomly chosen from the graph representing `karate` network, [12]. Left panel shows the number of entries of the vector $\mathbf{r}$ satisfying the condition $\mathbf{r}^{(t')} \leq \mathbf{r}^b$. As we see, at iteration $t = 9$ the vector $\mathbf{r}^b$ is reached and for any $t > 9$ the inequality is satisfied. The difference $\mathbf{r}^b - \mathbf{r}^{(t)}$ is shown on the right panel.

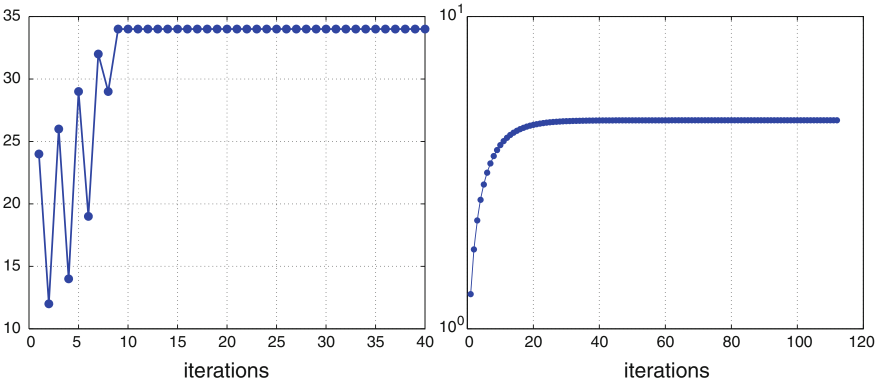


Fig. 4 Illustration of the fact that $\mathbf{r}^{(t')} \leq \mathbf{r}^b$ for any $t' > t$ (see text for the description)

ad.2. Assume the following value of the PageRank vector at time t:

$$r_j^f = \begin{cases} \frac{deg(j)}{\sum_{k \in U} deg(k)} & \text{if } j \in U \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

It is easily seen that the amount outflowing through each edge of each node in U amounts exactly to $\gamma = (1-\zeta)\frac{1}{\sum_{k \in U} deg(k)}$, while for the other nodes it is zero and $0 < \gamma$. Hence there exists in fact a configuration for which the starting conditions are matched.

Note also that such a reasoning supports our conviction that Eq. (4) can be true for $t = 0$ only. In subsequent steps the authority assigned to the nodes form the set U flows to the remaining nodes.

ad.3. Imagine now that instead of injecting into an empty network at the first moment the whole amount of authority $\mathbf{r}^f$, we inject at time $t = 0$ a fraction $\zeta \mathbf{r}^f$, at time $t = 1$ $(1-\zeta)\zeta \mathbf{r}^f$, and in general, at time $t = i$ a fraction $(1-\zeta)^i \zeta \mathbf{r}^f$. Then at moment $t = i$ in each edge no more authority than $\gamma \zeta \sum_{j=0}^{i} (1-\zeta)^i$ will flow so that in the limit γ will not be exceeded. But this is exactly what happens with an arbitrary initial vector $\mathbf{r}$ injected into the network at time t. Whatever its nature, ζ portion of it is injected via the supernode proportionally to $\mathbf{r}^f$ at time point 0, and so on, so that in the limit the initial authority vector plays no role at all. So the favorite situation will be reached in the end.

ad.4. Now it remains to show that the condition mentioned in step 1 of the proof really matches the thesis of the theorem.

Let us notice first that, due to the closed loop of authority circulation, the amount of authority flowing into U from the nodes belonging to the set $\overline{U} = N \backslash U$ must be identical with the amount flowing out of U to the nodes in $\overline{U}$.

But from U only that portion of authority flows out that flows out through the boundary of U because no authority leaves U via super-node (it returns from there immediately). As at most the amount $\gamma \partial(U)$ leaves U, then

$$p_o \zeta \leq \gamma \partial(U) = \frac{(1-\zeta)}{\sum_{k \in U} deg(k)} \partial(U) = (1-\zeta)\frac{\partial(U)}{Vol(U)} \qquad \square$$

Let us note in passing that $\frac{\partial(U)}{Vol(U)}$ is known as conductance in case that $Vol(U) \leq 0.5 Vol(N)$. Otherwise the denominator of conductance would be equal to $Vol(N) - Vol(U)$ (difference between the capacity of the whole network and the set U).

Observe also, that p_o, as the amount of the authority assigned to the nodes forming the set $\overline{U}$, can be expressed as $\mathbf{r}(\overline{U}) = 1 - \mathbf{r}(U)$. Denoting by $\mathfrak{c}(U)$ the conductance of U we obtain another form of the theorem

$$\mathbf{r}(U) \geq \frac{1-\zeta}{\zeta} \mathfrak{c}(U) \tag{5}$$

Finally, let us note that the assumption of the hub-oriented random walk is an important one because the above theorem is not valid for other choices of the distribution $\mathbf{s}$. To show this consider a numerical example. Let U has the form $U' \cup \{u\}$ where U' is a clique (complete sub-graph) consisting of 1000 nodes and u is a node having two neighbors: one belongs to the clique U' and the second is a node outside the clique. Assume further that only one node from U' has a link to a node outside the clique. Thus the volume of U is $Vol(U) = 1000 \times 999 + 1 + 2 =$999,003. Suppose that $\zeta = 1/2$. Would the theorem be true also for multipage personalized uniform random walker then one would have: $p_o \cdot 0.5 \leq (1 - 0.5)\frac{2}{999{,}003}$ which means $p_o \leq \frac{2}{999{,}003}$. But as each node of U gets from the super-node authority amounting to $\frac{0.5}{1.001}$, including the degree 2 node u, and thereafter, from this u node half of the authority would go outside of U, then $p_o \geq \frac{0.25}{1.001}$, which is an immediate contradiction.

3 Bipartite PageRank

Some non-directed graphs occurring e.g. in social networks are in a natural way bipartite graphs. Thus there exist nodes of two modalities and meaningful links may occur only between nodes of distinct modalities (e.g. authors and their papers, communities and their members, clients and items (or products) purchased by them, Fig. 5).

Some literature exists already for such networks attempting to adapt PageRank to the specific nature of bipartite graphs, e.g. [4]. Whatever investigations were performed, apparently no generalization of Theorem 1 was obtained.

One seemingly obvious choice would be to use the unimodal PageRank, like it was done in papers [1, 9], because a bipartite graph is in fact a graph. But this would

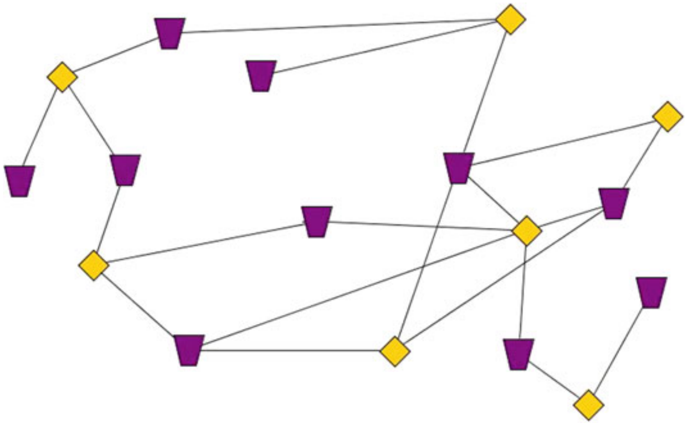


Fig. 5 An example of a bipartite graph

be conceptually wrong because the nature of the super-node would cause authority flowing between nodes of the same modality which is prohibited by the definition of these networks.

Therefore in this paper we intend to close this conceptual gap and will introduce a novel approach to Bipartite PageRank and will extend the Theorem 1 to this case.

So let us consider the flow of authority in a bipartite network with two distinct super-nodes: one collecting the authority from items and passing them to clients, and the other the authority from clients and passing them to items.

$$\mathbf{r}^p = (1 - \zeta^{cp}) \cdot \mathbf{P}^{cp} \cdot \mathbf{r}^c + \zeta^{cp} \cdot \mathbf{s}^p \tag{6}$$

$$\mathbf{r}^c = (1 - \zeta^{pc}) \cdot \mathbf{P}^{pc} \cdot \mathbf{r}^p + \zeta^{pc} \cdot \mathbf{s}^c \tag{7}$$

The following notation is used in these formulas

- $\mathbf{r}^p$, $\mathbf{r}^c$, $\mathbf{s}^p$, and $\mathbf{s}^c$ are stochastic vectors, i.e. the non-negative elements of these vectors sum to 1;
- the elements of matrix $\mathbf{P}^{cp}$ are: if there is a link from page j in the set of $Clients$ to a page i in the set of $Items$, then $p^{cp}_{ij} = \frac{1}{deg(j)}$, otherwise $p^{cp}_{ij} = 0$;
- the elements of matrix $\mathbf{P}^{pc}$ are: if there is a link from page j in the set of $Items$ to page i in the set of $Clients$, then $p^{pc}_{ij} = \frac{1}{deg(j)}$, and otherwise $p^{pc}_{ij} = 0$;
- $\zeta^{cp} \in [0, 1]$ is the boredom factor when jumping from *Clients* to *Items*;
- $\zeta^{pc} \in [0, 1]$ is the boredom factor when jumping from Items to Clients.

Definition 1 The solutions $\mathbf{r}^p$ and $\mathbf{r}^c$ of equation system (6) and (7) will be called item-oriented and client-oriented bipartite PageRanks, resp.

Again we assume that we have sets of preferred clients and items and we are interested in the outflow of authority towards the other items and clients.

Subsequently we will study first the flow of authority in some special cases to proceed to formulation of limiting theorems.

Let us assume first that

$$\zeta^{pc} = \zeta^{cp} = 0$$

i.e. that the super-nodes have no impact.

Let $K = \sum_{j \in Clients} deg(j) = \sum_{j \in Items} deg(j)$ mean the number of edges leaving one of the modalities. Let us consider a situation in which the amount of authority in each node $j \in Clients$ is $r^c_j = \frac{deg(j)}{K}$, and that $r^p_j = \frac{deg(j)}{K}$ for any $j \in Items$. Because through each link the same amount of $\frac{1}{K}$ authority will flow out, within each bidirectional link the amounts passed cancel out each other. So the $\mathbf{r}$'s defined this way are a fixed point (and solution) of the Eqs. (6) and (7).

For the other extreme, when $\zeta^{cp} = \zeta^{pc} = 1$ one obtains that $\mathbf{r}^p = \mathbf{s}^p$, $\mathbf{r}^c = \mathbf{s}^c$.

In analogy to the traditional PageRank let us note at this point that for $\zeta^{cp}, \zeta^{pc} > 0$ the "fan"-nodes of both the modalities (the sets of them being denoted with U^p for

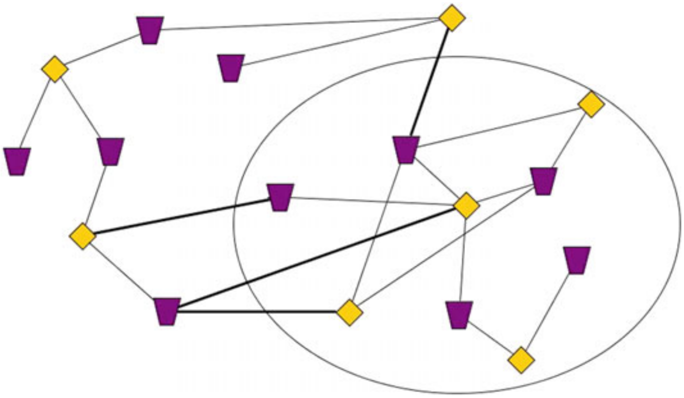


Fig. 6 Two sets of preferred nodes in a bipartite graph

items and U^c for clients), will obtain in each time step from the super-nodes the amount of authority equal to ζ^{pc} for clients and ζ^{pc} for products, resp.

Let us now think about a fan of the group of nodes U^p, U^c who prefers the hubs, and assume first that $U^p = Items, U^c = Clients$. Assume further that at the moment t we have the following state of authority distribution: node j contains $r_j^{c/p}(t) = \frac{deg(j)}{K}$ (meaning same formula for r^p and r^c). Let us consider now the moment $t+1$. From the product node j to the first super-node the authority $\zeta^{pc}\frac{deg(j)}{K}$ flows, and into each outgoing link $(1-\zeta^{pc})\frac{1}{K}$ is passed. On the other hand the client node j obtains from the same super-node authority $\zeta^{pc}\frac{indeg(j)}{K}$, while from each ingoing link $(1-\zeta^{pc})\frac{1}{K}$ what means that $\mathbf{r}^c(t+1) = \mathbf{r}^c(t)$. In an analogous way we can show that $\mathbf{r}^p(t+1) = \mathbf{r}^p(t)$. So a solution to the system of Eqs. (6) and (7) is found.

Let us now turn to the case when U^p, U^c are only proper subsets (or at least one of them) of *Items* and *Clients*, resp (Fig. 6). We shall be interested in the net amount of authority leaving each U^p, U^c to the nodes belonging to other modality via the outgoing links. This is equivalent to the amount of authority coming from the other nodes via supernodes to the sets U^p, U^c.

Applying techniques similar to the uni-modal case we can prove the theorem analogous to Theorem 1.

Theorem 2 *For the preferential personalized bipartite PageRank we have*

$$p_{c,o}\zeta^{cp} \leq (1-\zeta^{pc})\frac{|\partial(\frac{U^p}{U^c})|}{min(Vol(U^p), Vol(U^c))}$$

and

$$p_{p,o}\zeta^{pc} \leq (1-\zeta^{cp})\frac{|\partial(\frac{U^c}{U^p})|}{min(Vol(U^p), Vol(U^c))}$$

where

- $p_{c,o}$ *is the sum of authorities from the set* $Clients \backslash U^c$*,*
- $p_{p,o}$ *is the sum of authorities from the set* $Items \backslash U^p$*,*
- $\partial(\frac{U^c}{U^p})$ *is the set of edges outgoing from* U_c *into nodes from* $Items \backslash U^p$ *(that is "fan's border" of* U^c*),*
- $\partial(\frac{U^p}{U^c})$ *is the set of edges outgoing from* U^p *into nodes from* $Clients \backslash U^c$ *(that is "fan's border" of* U^p*),*
- $Vol(U^c)$ *is the sum of degrees of all nodes from* U^c *(capacity of* U^c*)*
- $Vol(U^p)$ *is the sum of degrees of all nodes from* U^p *(capacity of* U^p*).* □

Proof We have a painful surprise this time. In general we cannot define a useful state of authority of nodes, analogous to that of traditional PageRank from the previous section, so that in both directions between U^p and U^c nodes the same amount of authority would flow. This is due to the fact that in general capacities of U^c and U^p may differ. Therefore a broader generalization is required.

To find such a generalization let us reconsider the way how we can limit the flow of authority in a single "channel" (link from one node to the other). The amount of authority passed consists of two parts: a variable one being a share of the authority at the feeding end of the channel and a fixed one coming from a super-node. So, by increasing the variable part we come to the point that the receiving end gets less authority than this node fed into the channel—this is because of increased "evaporation" to the supernode.

Let us seek the amount of authority d being an upper bound on authority residing in a node per its incidental link such that if this limitation is obeyed in a given step, it is also obeyed in the subsequent ones. So let i be a client node of degree $deg(i)$ and j be a product node of degree $deg(j)$. This means that the amount of authority at i shall not exceed $d \cdot deg(i)$ and $d \cdot deg(j)$ at j. In the next step node i gets the authority of no more than $\zeta^{pc}\frac{deg(i)}{\sum_{v\in U^c} deg(v)} + deg(i)\cdot d \cdot (1-\zeta^{pc})$. Node j gets the authority of no more than $\zeta^{cp}\frac{deg(j)}{\sum_{v\in U^p} deg(v)} + deg(j)\cdot d\cdot(1-\zeta^{cp})$. We are interested if the following holds:

$$d \cdot deg(i) \geq \zeta^{pc}\frac{deg(i)}{\sum_{v\in U^c} deg(v)} + deg(i)\cdot d\cdot(1-\zeta^{pc})$$

$$d \cdot deg(j) \geq \zeta^{cp}\frac{deg(j)}{\sum_{v\in U^p} deg(v)} + deg(j)\cdot d\cdot(1-\zeta^{cp})$$

This may be simplified to

$$d \geq \zeta^{pc}\frac{1}{\sum_{v\in U^c} deg(v)} + d\cdot(1-\zeta^{pc})$$

$$d \geq \zeta^{cp} \frac{1}{\sum_{v \in U^p} deg(v)} + d \cdot (1 - \zeta^{cp})$$

Hence:

$$d \cdot \zeta^{pc} \geq \zeta^{pc} \frac{1}{\sum_{v \in U^c} deg(v)}, \quad d \cdot \zeta^{cp} \geq \zeta^{cp} \frac{1}{\sum_{v \in U^p} deg(v)}$$

And finally:

$$d \geq \frac{1}{\sum_{v \in U^c} deg(v)}, \quad d \geq \frac{1}{\sum_{v \in U^p} deg(v)}$$

So if we denote $Vol(U^p) = \sum_{v \in U^p} deg(v)$ and $Vol(U^c) = \sum_{v \in U^c} deg(v)$ we get finally

$$d = max \left(\frac{1}{Vol(U^p)}, \frac{1}{Vol(U^c)} \right) = \frac{1}{min(Vol(U^p), Vol(U^c))}$$

Let us notice first that, due to the closed loop of authority circulation, the amount of authority flowing into U^c from the nodes belonging to the set $\overline{U^p} = Items \backslash U^p$ must be identical with the amount flowing out of U^p to the nodes in $\overline{U^c}$. The same holds when we interchange the indices p and c.

But from U^p only that portion of authority flows out to $\overline{U^c}$ that flows out through the boundary of U^p because no authority leaves the tandem U^p, U^c via super-nodes (it returns from there immediately). As the amount $d(1 - \zeta^{pc})\partial(\frac{U^p}{U^c})$ leaves at most the U^p through outlinks not going into U^c, then

$$p_{c,o}\zeta^{cp} \leq d(1 - \zeta^{pc}) \left| \partial(\frac{U^p}{U^c}) \right| = (1 - \zeta^{pc}) \frac{1}{min(Vol(U^p), Vol(U^c))} \left| \partial(\frac{U^p}{U^c}) \right|$$

For the flow from U^c to $\overline{U^p}$ we have analogous derivation.

Let us now consider the issue if the flow d is obtainable in the links by specially initializing the authorities of the nodes of the network. Assume that each node j of U^p was initialized to $\frac{deg(j)}{Vol(U^p)}$ and each node i of U^c was initialized to $\frac{deg(i)}{Vol(U^c)}$ and the remaining nodes are zero which is a valid initialization as the sum over nodes in each modality is equal 1. Under these circumstances the authority of each node per link is $\frac{1}{Vol(U^c)}$ for U^c nodes and $\frac{1}{Vol(U^p)}$ for U^p nodes which is precisely what we need.

The last part of the proof is to show that this advantegous condition will be achieved in the limit and here we just can repeat the respective fragment of the proof of the preceding theorem. Again of course ζ's must be non-zero. □

Note by the way that the convergence is achieved in an analogous way as done for the HITS (consult e.g. [8, Chap. 11]).

4 Concluding Remarks

In this paper we have introduced a novel approach to the concept of bipartite Page Rank and have proposed limits for the flow of authority in a bipartite graph.

These limits can be used for example when verifying validity of clusters in such graphs. It is quite common to assume that the better the cluster the less authority flows out of it when treating the cluster as the set on which a fan concentrates while a personalized PageRank is computed. The theorem says that the outgoing authority has a natural upper limit dropping with the growth of the size of the sub-network so that the outgoing authority cluster validity criterion cannot be used because it will generate meaningless large clusters. So a proper validity criterion should make a correction related to the established limits in order to be of practical use.

In fact, this should be combined with a bootstrapping approach: an outgoing authority should be compared with randomly generated structures of the cluster network to find out if this outgoing authority for a given structure is likely to differ from a random structure.

To justify this remark note that the primary state-of-the-art mistake of the state-of-the-art-clustering algorithms is the lack of any verification if the discovered structure of clusters is really a new piece of knowledge or just an artifact of large scaled random processes. This affects also the numerous graph clustering algorithms. Therefore we propose in particular the following:

(1) significance of the clusters should be tested, and
(2) quality of the clusters should be tested.

The significance should be tested using Monte Carlo method as follows: For a cluster, count the number of edges in it. Next create random graphs over the set of cluster nodes of the same size in terms of the number of edges, according to some random process (requesting e.g. connectivity). For the original graph, substitute the cluster with the newly generated random graph and compute the personalized PageRank. Repeat this a number of times and count the share of random graphs for which the amount of out-flowing authority from the cluster is lower than for the original graph. If it is lower than say 5 %, the cluster is significant.

For significant clusters compute the quotient of outgoing authority with the theoretical limits. The best clusters are those with the lowest quotient.

As a further research direction it is obvious that finding tighter limits is needed, or a proof that the found limits are the lowest ones possible. This would improve the evaluation of e.g. cluster quality.

Acknowledgments This research has been supported by the Polish State budget scientific research funds.

References

1. Bauckhage C (2008) Image tagging using pagerank over bipartite graphs. In: Proceedings of the 30th DAGM symposium on pattern recognition, pp 426–435. Springer, Berlin, Heidelberg. doi:10.1007/978-3-540-69321-5_43
2. Berkhin P (2005) A survey on PageRank computing. Internet Math 2:73–120
3. Chung F (2011) Pagerank as a discrete Green's function. In: Ji L (ed) Geometry and analysis, I, Advanced lectures in mathematics (ALM), vol 17, pp 285–302. International Press of Boston, Boston
4. Deng H, Lyu MR, King I (2009) A generalized co-hits algorithm and its application to bipartite graphs. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'09, pp 239–248. ACM, New York, NY, USA, Paris, June 28-July 1 2009 doi:10.1145/1557019.1557051
5. Frahm K, Georgeot B, Shepelyansky D (2011) Universal emergence of PageRank. J Phys A Math Theor 44:465101. doi:10.1088/1751-8113/44/46/465101
6. Garcia E, Pedroche F, Romance M (2013) On the localization of the personalized PageRank of complex networks. Linear Algebra Appl 439:640–652
7. Langville AN (2005) An annotated bibliography of papers about Markov chains and information retrieval. http://www.cofc.edu/langvillea/bibtexpractice.pdf
8. Langville AN, Meyer CD (2006) Google's PageRank and beyond: the science of search engine rankings. Princeton University Press, Princeton
9. Link S (2011) Eigenvalue-based bipartite ranking. Bachelorarbeit/bachelor thesis. http://www.pms.ifi.lmu.de/publikationen/#PA_Stephan.Link
10. Meghabghab G, Kandel A (2008) Search engines, link analysis, and user's web behavior. A unifying web mining approach, Studies in computational intelligence, vol 99. Springer, New York
11. Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. Technical Report 1999–66, Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/
12. Zachary W (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33:452–473

Dependence Factor as a Rule Evaluation Measure

Marzena Kryszkiewicz

Abstract Certainty factor and lift are known evaluation measures of association rules. Nevertheless, they do not guarantee accurate evaluation of the strength of dependence between rule's constituents. In particular, even if there is a strongest possible positive or negative dependence between rule's constituents X and Y, these measures may reach values quite close to the values indicating independence of X and Y. Recently, we have proposed a new measure called a dependence factor to overcome this drawback. Unlike in the case of the certainty factor, when defining the dependence factor, we took into account the fact that for a given rule $X \rightarrow Y$, the minimal conditional probability of the occurrence of Y given X may be greater than 0, while its maximal possible value may less than 1. In this paper, we first recall definitions and properties of all the three measures. Then, we examine the dependence factor from the point of view of an interestingness measure as well as we examine the relationship among the dependence factor for X and Y with those for $\bar{X}$ and Y, X and $\bar{Y}$, as well as $\bar{X}$ and $\bar{Y}$, respectively. As a result, we obtain a number of new properties of the dependence factor.

1 Introduction

Certainty factor and *lift* are known measures of association rules. The former measure was offered in the expert system Mycin [9], while the latter is widely implemented in both commercial and non-commercial data mining systems [2]. Nevertheless, they do not guarantee accurate evaluation of the strength of dependence between rule's constituents. In particular, even if there is a strongest possible positive or negative dependence between rule's constituents X and Y, these measures may reach values quite close to the values indicating independence of X and Y. This might suggest that one deals with a weak dependence, while in fact the dependence is strong. In [4],

M. Kryszkiewicz (✉)
Institute of Computer Science, Warsaw University of Technology,
Nowowiejska 15/19, 00-665 Warsaw, Poland
e-mail: mkr@ii.pw.edu.pl

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_12

we proposed a new measure called a *dependence factor* to overcome this drawback. Unlike in the case of the certainty factor, when defining the dependence factor, we took into account the fact that for a given rule $X \rightarrow Y$, the minimal conditional probability of the occurrence of Y given X may be greater than 0, while its maximal possible value may less than 1. The dependence factor always takes value 1 if a dependence is strongest possible positive, whereas for a strongest possible negative dependence, it always takes value –1; in the case of independence, it takes value 0.

In [4], we have focused on examining properties of the dependence factor as a measure of dependence between rule's constituents/events. Our new main contribution in this paper is: (1) the examination of the dependence factor as an interestingness measure with respect to the interestingness postulates formulated by Piatetsky-Shapiro in [7], and (2) the derivation of the relationship among the dependence factor for X and Y, with those for $\bar{X}$ and Y, X and $\bar{Y}$, as well as $\bar{X}$ and $\bar{Y}$, respectively.

Our paper has the following layout. In Sect. 2, we briefly recall basic notions of association rules, their basic measures (support, confidence) as well as lift and certainty factor. In Sect. 3, we recall maximal and minimal values of examined measures in the case when probabilities of rule's constituents are fixed, as shown in [4]. In Sect. 4, we recall the definition and properties of the dependence factor after [4]. Our new contribution is presented in Sects. 5 and 6. In Sect. 5, we examine the usefulness of the dependence factor as an interestingness measure, while in Sect. 6, we identify the relationship between the dependence factors for events and their complements. Section 7 concludes our work.

2 Basic Notions and Properties

In this section, we recall the notion of association rules after [1].

Definition 1 Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of distinct literals, called *items* (e.g. products, features, symptoms). Any subset X of the set I is called an *itemset*. A *transaction database* is denoted by $\mathcal{D}$ and is defined as a set of itemsets. Each itemset T in $\mathcal{D}$ is a *transaction*. An *association rule* is an expression associating two itemsets:

$$X \rightarrow Y, \text{ where } \emptyset \neq Y \subseteq I \text{ and } X \subseteq I \backslash Y.$$

Itemsets and association rules are typically characterized by *support* and *confidence*, which are simple statistical parameters.

Definition 2 *Support* of an itemset X is denoted by $sup(X)$ and is defined as the number of transactions in $\mathcal{D}$ that contain X; that is

$$sup(X) = |\{T \in \mathcal{D} | X \subseteq T\}|.$$

Support of a rule $X \rightarrow Y$ is denoted by $sup(X \rightarrow Y)$ and is defined as the support of $X \cup Y$; that is,

$$sup(X \rightarrow Y) = sup(X \cup Y).$$

Clearly, the probability of the event that itemset X occurs in a transaction equals $sup(X)/|\mathscr{D}|$, while the probability of the event that both X and Y occur in a transaction equals $sup(X \cup Y)/|\mathrm{D}|$. In the remainder, the former probability will be denoted by $P(X)$, while the latter by $P(XY)$.

Definition 3 The *confidence* of an association rule $X \rightarrow Y$ is denoted by $conf(X \rightarrow Y)$ and is defined as the conditional probability that Y occurs in a transaction provided X occurs in the transaction; that is:

$$conf(X \rightarrow Y) = \frac{sup(X \rightarrow Y)}{sup(X)} = \frac{P(XY)}{P(X)}.$$

A large amount of research was devoted to *strong association rules* understood as those association rules the supports and confidences of which exceed user-defined support threshold and confidence threshold, respectively. However, it has been argued in the literature that these two measures are not sufficient to express different interestingness, usefulness or unexpectedness aspects of association rules [3, 5–8, 10–12]. In fact, a number of such measures of association rules was proposed (see e.g. [3, 5–8, 10–12]. Among them very popular measures are *lift* [2] and *certainty factor* [9].

Definition 4 The *lift* of an association rule $X \rightarrow Y$ is denoted by $lift(X \rightarrow Y)$ and is defined as the ratio of the conditional probability of the occurrence of Y in a transaction given X occurs there to the probability of the occurrence of Y; that is:

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{P(Y)}.$$

Lift may be also defined in an equivalent way in terms of probabilities only:

Property 1

$$lift(X \rightarrow Y) = \frac{P(XY)}{P(X) \times P(Y)}.$$

Definition 5 The *certainty factor* of an association rule $X \rightarrow Y$ is denoted by $cf(X \rightarrow Y)$ and is defined as the degree to which the probability of the occurrence of Y in a transaction can change when X occurs there as follows:

$$cf(X \rightarrow Y) = \begin{cases} \frac{conf(X \rightarrow Y) - P(Y)}{1 - P(Y)} & \text{if } conf(X \rightarrow Y) > P(Y), \\ 0 & \text{if } conf(X \rightarrow Y) = P(Y), \\ -\frac{P(Y) - conf(X \rightarrow Y)}{P(Y) - 0} & \text{if } conf(X \rightarrow Y) < P(Y). \end{cases}$$

The definition of the certainty factor is based on the assumption that the probability of the occurrence of Y in a transaction given X occurs there ($conf(X \rightarrow Y)$) can

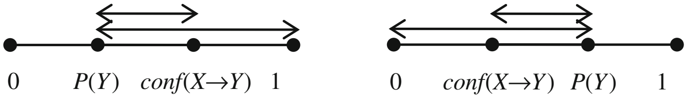


Fig. 1 Calculating the absolute value of the certainty factor as the ratio of the lengths of respective intervals when $conf(X \rightarrow Y) > P(Y)$ (on the *left-hand side*) and when $conf(X \rightarrow Y) < P(Y)$ (on the *right-hand side*)

be increased from $P(Y)$ up to 1 and decreased from $P(Y)$ down to 0. In Fig. 1, we visualize the meaning of the absolute value of the certainty factor as the ratio of the lengths of respective intervals.

As shown in Property 2, the certainty factor can be expressed equivalently in terms of unconditional probabilities (by multiplying the numerator and denominator of the formula in Definition 5 by $P(X)$) or lift (by dividing the numerator and denominator of the original *cf* formula by $P(Y)$).

Property 2

$$(a)\ \ cf(X \rightarrow Y) = \begin{cases} \frac{P(XY) - P(X) \times P(Y)}{P(X) - P(X) \times P(Y)} & \text{if } P(XY) > P(X) \times P(Y), \\ 0 & \text{if } P(XY) = P(X) \times P(Y), \\ -\frac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - 0} & \text{if } P(XY) < P(X) \times P(Y). \end{cases}$$

$$(b)\ \ cf(X \rightarrow Y) = \begin{cases} \frac{lift(X \rightarrow Y) - 1}{\frac{1}{P(Y)} - 1} & \text{if } lift(X \rightarrow Y) > 1, \\ 0 & \text{if } lift(X \rightarrow Y) = 1, \\ -\frac{1 - lift(X \rightarrow Y)}{1 - 0} & \text{if } lift(X \rightarrow Y) < 1. \end{cases}$$

Both lift and certainty factor are related to the notion of (in)dependence of events, where two events are treated as independent if the product of the probabilities of their occurrences equals the probability that the two events co-occur. Otherwise, they are regarded as dependent. Note that this notion of dependence does not indicate which event is a reason of the other. However, it allows formulating whether the dependence between events is positive or negative in the case when the events are dependent on each other.

Definition 6 *X* and *Y* are:

- *independent* if $P(XY) = P(X) \times P(Y)$,
- *dependent positively* if $P(XY) > P(X) \times P(Y)$,
- *dependent negatively* if $P(XY) < P(X) \times P(Y)$.

In Table 1, we provide equivalent conditions in terms of *P*, *conf*, *lift* and *cf* for independence, positive dependence and negative dependence, respectively, between two itemsets.

In general, one may distinguish between *symmetric* (*two direction*) *measures* of association rules and *asymmetric* (*one direction*) ones.

Table 1 Conditions for independence, positive dependence and negative dependence

(In)dependence	(In)dependence condition	Equivalent conditions in terms of measures for $X \rightarrow Y$	Equivalent conditions in terms of measures for $Y \rightarrow X$
Y and X are dependent positively	$P(XY) > P(X) \times P(Y)$	$conf(X \rightarrow Y) > P(Y)$ $lift(X \rightarrow Y) > 1$ $cf(X \rightarrow Y) > 0$	$conf(Y \rightarrow X) > P(X)$ $lift(Y \rightarrow X) > 1$ $cf(Y \rightarrow X) > 0$
Y and $X$are independent	$P(XY) = P(X) \times P(Y)$	$conf(X \rightarrow Y) = P(Y)$ $lift(X \rightarrow Y) = 1$ $cf(X \rightarrow Y) = 0$	$conf(Y \rightarrow X) = P(X)$ $lift(Y \rightarrow X) = 1$ $cf(Y \rightarrow X) = 0$
Y and $X$are dependent negatively	$P(XY) < P(X) \times P(Y)$	$conf(X \rightarrow Y) < P(Y)$ $lift(X \rightarrow Y) < 1$ $cf(X \rightarrow Y) < 0$	$conf(Y \rightarrow X) < P(X)$ $lift(Y \rightarrow X) < 1$ $cf(Y \rightarrow X) < 0$

Definition 7 A measure m is called *symmetric* (*two direction*) if $m(X \rightarrow Y) = m(Y \rightarrow X)$ for any X and Y. Otherwise, it is called an *asymmetric* (*one direction*) *measure*.

Property 3

(a) $conf(X \rightarrow Y) = conf(Y \rightarrow X)$ *is not guaranteed to hold.*
(b) $lift(X \rightarrow Y) = lift(Y \rightarrow X)$.
(c) $cf(X \rightarrow Y) = cf(Y \rightarrow X)$ *is not guaranteed to hold if* $conf(X \rightarrow Y) > P(Y)$.
(d) $cf(X \rightarrow Y) = cf(Y \rightarrow X)$ *if* $conf(X \rightarrow Y) \leq P(Y)$.

As follows from Property 3, *conf* is an asymmetric measure and *lift* is a symmetric measure. On the other hand, we observe that strangely *cf* has a mixed nature—asymmetric for positive dependences and symmetric for negative dependences and independences. This observation provoked us to revisit the definition of *cf* and to propose its modification in [4]. When defining the dependence factor there, we took into account the fact that in some circumstances it may be infeasible to increase the probability of the occurrence of Y in a transaction under the presence of X ($conf(X \rightarrow Y)$) from $P(Y)$ up to 1 as well as it may be infeasible to decrease it from $P(Y)$ down to 0.

3 Maximal and Minimal Values of Rule Measures

In this section, we first recall global maximal and minimal values of rule measures (Table 2). Next, following [4], we recall maximal and minimal values of rule measures for given values of $P(X)$ and $P(Y)$.

In the remainder of the paper, we denote *maximal probability* and *minimal probability* of the co-occurrence of X and Y given $P(X)$ and $P(Y)$ are fixed by $max\_P(XY|_{P(X),P(Y)})$ and $min\_P(XY|_{P(X),P(Y)})$, respectively. Analogously, *maximal confidence* and *minimal*

Table 2 Global maximal and minimal values of rule measures

Measure	Max	Min
$P(XY)$	1	0
$conf(X \to Y)$	1	0
$lift(X \to Y)$	∞	0
$cf(X \to Y)$	1 if Y depends on X positively	-1 if Y depends on X negatively

confidence (*maximal lift*, *minimal lift*, *maximal certainty factor*, *minimal certainty factor*) *of* $X \to Y$ given $P(X)$ and $P(Y)$ are fixed are denoted by $max\_conf(X \to Y|_{P(X),P(Y)})$ and $min\_conf(X \to Y|_{P(X),P(Y)})$ ($max\_lift(X \to Y|_{P(X),P(Y)})$, $min\_lift(X \to Y|_{P(X),P(Y)})$, $max\_cf(X \to Y|_{P(X),P(Y)})$, $min\_cf(X \to Y|_{P(X),P(Y)})$, respectively.

Property 4

$$(a)\ \ max\_conf(X \to Y|_{P(X),P(Y)}) = \frac{max\_P(XY|_{P(X),P(Y)})}{P(X)}$$

$$(b)\ \ min\_conf(X \to Y|_{P(X),P(Y)}) = \frac{min\_P(XY|_{P(X),P(Y)})}{P(X)}$$

$$(c)\ \ max\_lift(X \to Y|_{P(X),P(Y)}) = \frac{max\_conf(XY|_{P(X),P(Y)})}{P(Y)} = \frac{max\_P(XY|_{P(X),P(Y)})}{P(X) \times P(Y)}$$

$$(d)\ \ min\_lift(X \to Y|_{P(X),P(Y)}) = \frac{min\_conf(XY|_{P(X),P(Y)})}{P(Y)} = \frac{min\_P(XY|_{P(X),P(Y)})}{P(X) \times P(Y)}$$

$$(e)\ \ max\_cf(X \to Y|_{P(X),P(Y)}) = \frac{max\_conf(X \to Y|_{P(X),P(Y)}) - P(Y)}{1 - P(Y)}$$
$$= \frac{max\_P(XY|_{P(X),P(Y)}) - P(X) \times P(Y)}{P(X) - P(X) \times P(Y)} = \frac{max\_lift(XY|_{P(X),P(Y)}) - 1}{\frac{1}{P(Y)} - 1}$$

$$(f)\ \ min\_cf(X \to Y|_{P(X),P(Y)}) = -\frac{P(Y) - min\_conf(X \to Y|_{P(X),P(Y)})}{P(Y) - 0}$$
$$= -\frac{P(X) \times P(Y) - min\_P(XY|_{P(X),P(Y)})}{P(X) \times P(Y) - 0} = -\frac{1 - min\_lift(XY|_{P(X),P(Y)})}{1 - 0}$$

In Proposition 1, we show how to calculate $min\_P(XY|_{P(X),P(Y)})$ and $max\_P(XY|_{P(X),P(Y)})$. We note that neither $max\_P(XY|_{P(X),P(Y)})$ necessarily equals 1 nor $min\_P(XY|_{P(X),P(Y)})$ necessarily equals 0. Figure 2 illustrates this.

Proposition 1

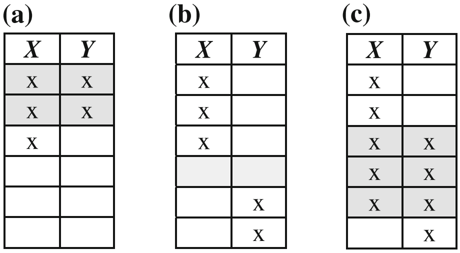
$$(a)\ \ max\_P(XY|_{P(X),P(Y)}) = \min\{P(X), P(Y)\}$$

$$(b)\ \ min\_P(XY|_{P(X),P(Y)}) = \begin{cases} 0 & \text{if } P(X) + P(Y) \le 1 \\ P(X) + P(Y) - 1 & \text{if } P(X) + P(Y) > 1 \end{cases}$$
$$= \max\{0, P(X) + P(Y) - 1\}$$

The next proposition follows from Property 4 and Proposition 1.

Proposition 2

$$(a)\ \ max\_conf(X \to Y|_{P(X),P(Y)}) = \frac{\min\{P(X), P(Y)\}}{P(X)} =$$
$$\begin{cases} 1 & \text{if } P(X) \le P(Y), \\ \frac{P(Y)}{P(X)} & \text{if } P(Y) < P(X). \end{cases}$$



(a)

X	*Y*
x	x
x	x
x	

(b)

X	*Y*
x	
x	
x	
	x
	x

(c)

X	*Y*
x	
x	
x	x
x	x
x	x
	x

Fig. 2 **a** $max\_P(XY|_{P(X),P(Y)}) = \min\{P(X), P(Y)\} = \min\left\{\frac{3}{6}, \frac{2}{6}\right\} = \frac{2}{6}$. **b** $min\_P(XY|_{P(X),P(Y)}) = 0$ if $P(X) + P(Y) \leq 1$. **c** $min\_P(XY|_{P(X),P(Y)}) = P(X) + P(Y) - 1 = \frac{5}{6} + \frac{4}{6} - 1 = \frac{3}{6}$ if $P(X) + P(Y) > 1$

$$(b)\ \ min\_conf(X \rightarrow Y|_{P(X),P(Y)}) = \frac{\max\{0, P(X)+P(Y)-1\}}{P(X)}$$

$$= \begin{cases} 0 & \text{if } P(X) + P(Y) \leq 1, \\ \frac{P(X)+P(Y)-1}{P(X)} & \text{if } P(X) + P(Y) > 1. \end{cases}$$

$$(c)\ \ max\_lift(X \rightarrow Y|_{P(X),P(Y)}) = \frac{\min\{P(X),P(Y)\}}{P(X) \times P(Y)} = \frac{1}{max\{P(X),P(Y)\}}.$$

$$(d)\ \ min\_lift(X \rightarrow Y|_{P(X),P(Y)}) = \frac{\max\{0, P(X)+P(Y)-1\}}{P(X) \times P(Y)}$$

$$= \begin{cases} 0 & \text{if } P(X) + P(Y) \leq 1, \\ \frac{P(X)+P(Y)-1}{P(X) \times P(Y)} & \text{if } P(X) + P(Y) > 1. \end{cases}$$

$$(e)\ \ max\_cf(X \rightarrow Y|_{P(X),P(Y)}) = \frac{\min\{P(X),P(Y)\} - P(X) \times P(Y)}{P(X) - P(X) \times P(Y)}$$

$$= \frac{\frac{1}{\max\{P(X),P(Y)\}} - 1}{\frac{1}{P(Y)} - 1} = \begin{cases} 1 & \text{if } P(X) \leq P(Y), \\ \frac{\frac{1}{P(X)} - 1}{\frac{1}{P(Y)} - 1} & \text{if } P(X) > P(Y). \end{cases}$$

$$(f)\ \ min\_cf(X \rightarrow Y|_{P(X),P(Y)}) = -\frac{P(X) \times P(Y) - \max\{0, P(X)+P(Y)-1\}}{P(X) \times P(Y) - 0}$$

$$= \frac{\max\{0, P(X)+P(Y)-1\}}{P(X) \times P(Y)} - 1 = \begin{cases} -1 & \text{if } P(X) + P(Y) \leq 1 \\ \frac{P(X)+P(Y)-1}{P(X) \times P(Y)} - 1 & \text{if } P(X) + P(Y) > 1. \end{cases}$$

In Table 3, we summarize real achievable maximal and minimal values of $P(XY)$, $conf(X \rightarrow Y)$, $lift(X \rightarrow Y)$ and $cf(X \rightarrow Y)$ for given values of $P(X)$ and $P(Y)$.

4 Dependence Factor

In this section, we recall the definition of the *dependence factor* of a rule $X \rightarrow Y$, which we offered in [4] as a modification of the certainty factor. Unlike the certainty factor, it is based on real maximal and minimal values of $conf(X \rightarrow Y)$ for given values of $P(X)$ and $P(Y)$. Then we present the properties of this measure.

Table 3 Real achievable maximal and minimal values of $P(XY)$, $conf(X \rightarrow Y)$, $lift(X \rightarrow Y)$ and $cf(X \rightarrow Y)$ for given values of $P(X)$ and $P(Y)$

Measure	Max for given values of $P(X)$ and $P(Y)$	Min for given values of $P(X)$ and $P(Y)$
$P(XY)$	$\min\{P(X), P(Y)\}$	$\max\{0, P(X) + P(Y) - 1\}$
$conf(X \rightarrow Y)$	$\frac{\min\{P(X), P(Y)\}}{P(X)}$	$\frac{\max\{0, P(X)+P(Y)-1\}}{P(X)}$
$lift(X \rightarrow Y)$	$\frac{\min\{P(X), P(Y)\}}{P(X) \times P(Y)}$	$\frac{\max\{0, P(X)+P(Y)-1\}}{P(X) \times P(Y)}$
$cf(X \rightarrow Y)$	$\frac{\min\{P(X), P(Y)\}-P(X) \times P(Y)}{P(X)-P(X) \times P(Y)}$ if Y depends on X positively	$-\frac{P(X) \times P(Y)-\max\{0, P(X)+P(Y)-1\}}{P(X) \times P(Y)-0}$ if Y depends on X negatively

Definition 8 The *dependence factor* of $X \rightarrow Y$ is denoted by $df(X \rightarrow Y)$ and is defined as the ratio of the actual change of the probability of the occurrence of Y in a transaction given X occurs there to its maximal feasible change as follows:

$$df(X \rightarrow Y) = \begin{cases} \frac{conf(X \rightarrow Y) - P(Y)}{max\_conf(X \rightarrow Y|\ P(X),\ P(Y)) - P(Y)} & \text{if } conf(X \rightarrow Y) > P(Y), \\ 0 & \text{if } conf(X \rightarrow Y) = P(Y), \\ -\frac{P(Y) - conf(X \rightarrow Y)}{P(Y) - min\_conf(X \rightarrow Y|_{P(X),\ P(Y)})} & \text{if } conf(X \rightarrow Y) < P(Y). \end{cases}$$

The dependence factor not only determines by how much the probability of the occurrence of Y in a transaction changes under the presence of X with respect to by how much it could have changed, but also it determines by how much the probability of the occurrence of X and Y in a transaction differs from the probability of their common occurrence under independence assumption with respect to by how much it could have been different (see Proposition 3a). In addition, the dependence factor determines by how much the value of the lift of a rule $X \rightarrow Y$ differs from the value 1 (that is, from the value indicating independence of rule's constituents in terms of the lift measure) with respect to by how much it could have been be different (see Proposition 3b).

Proposition 3

$$(a)\quad df(X \rightarrow Y) = \begin{cases} \frac{P(XY) - P(X) \times P(Y)}{max\_P(XY|_{P(X),P(Y)}) - P(X) \times P(Y)} & \text{if } P(XY) > P(X) \times P(Y), \\ 0 & \text{if } P(XY) = P(X) \times P(Y), \\ -\frac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - min\_P(XY|_{P(X),P(Y)})} & \text{if } P(XY) < P(X) \times P(Y). \end{cases}$$

$$(b)\quad df(X \rightarrow Y) = \begin{cases} \frac{lift(X \rightarrow Y) - 1}{max\_lift(X \rightarrow Y|_{P(X),P(Y)}) - 1} & \text{if } lift(X \rightarrow Y) > 1, \\ 0 & \text{if } lift(X \rightarrow Y) = 1, \\ -\frac{1 - lift(X \rightarrow Y)}{1 - min\_lift(X \rightarrow Y|_{P(X),P(Y)})} & \text{if } lift(X \rightarrow Y) < 1. \end{cases}$$

Theorem 1

(a) If $P(XY) > P(X) \times P(Y)$, then $df(X \rightarrow Y) \in (0, 1]$.
(b) If $P(XY) = P(X) \times P(Y)$, then $df(X \rightarrow Y) = 0$.
(c) If $P(XY) < P(X) \times P(Y)$, then $df(X \rightarrow Y) \in [-1, 0)$.

Proof Follows from Proposition 3a. □

Table 4 Maximal and minimal values of $df(X \to Y)$ for any given values of $P(X)$ and $P(Y)$

Measure	Max for any given values of $P(X)$ and $P(Y)$	Min for any given values of $P(X)$ and $P(Y)$
$df(X \to Y)$	1 if X and Y are dependent positively	–1 if X and Y are dependent negatively

As follows from Proposition 3a, the dependence factor is a symmetric measure.

Theorem 2 $df(X \to Y) = df(Y \to X)$.

Based on Proposition 1 and 3a, we will express the dependence factor $df(X \to Y)$ in terms of $P(XY)$, $P(X)$ and $P(Y)$, which will be useful for examining properties of this measure.

Theorem 3

$$df(X \to Y) = \begin{cases} \frac{P(XY) - P(X) \times P(Y)}{\min\{P(X), P(Y)\} - P(X) \times P(Y)} & \text{if } P(XY) > P(X) \times P(Y), \\ 0 & \text{if } P(XY) = P(X) \times P(Y), \\ -\frac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - \max\{0, P(X) + P(Y) - 1\}} & \text{if } P(XY) < P(X) \times P(Y). \end{cases}$$

One may easily note that $df(X \to Y)$ reaches 1 when $P(XY)$ is maximal for given values of $P(X)$ and $P(Y)$; that is, when $P(XY) = \min\{P(X), P(Y)\}$ or, in other words, when the dependence between X and Y is strongest possible positive for given values of $P(X)$ and $P(Y)$. Analogously, $df(X \to Y)$ reaches –1 when $P(XY)$ is minimal for given values of $P(X)$ and $P(Y)$; that is, when $P(XY) = \max\{0, P(X) + P(Y) - 1\}$ or, in other words, when the dependence between X and Y is strongest possible negative for these probability values (Table 4).

Based on Theorem 3 and Property 2a, one may derive relations between the dependence factor and the certainty factor as follows:

Theorem 4

(a) $df(X \to Y) \geq cf(X \to Y)$ if $P(XY) > P(X) \times P(Y)$,
(b) $df(X \to Y) = cf(X \to Y) = 0$ if $P(XY) = P(X) \times P(Y)$,
(c) $df(X \to Y) \leq cf(X \to Y)$ if $P(XY) < P(X) \times P(Y)$,
(d) $df(X \to Y) = \max\{cf(X \to Y), cf(Y \to X)\}$ if $P(XY) > P(X) \times P(Y)$,
(e) $df(X \to Y) = cf(X \to Y)$ if $P(XY) < P(X) \times P(Y)$ *and* $P(X) + P(Y) < 1$,
(f) $df(X \to Y) < cf(X \to Y)$ if $P(XY) < P(X) \times P(Y)$ *and* $P(X) + P(Y) > 1$.

Tables 5–6 illustrate the findings expressed as Theorem 4. In particular, Table 5 shows values of $lift(X \to Y)$, $cf(X \to Y)$ and $df(X \to Y)$ for $P(X) = 0.6$ and $P(Y) = 0.3$; that is, in the case when $P(X) + P(Y) \leq 1$. For these values of $P(X)$ and $P(Y)$, the maximal possible value for $P(XY)$ equals $\min\{P(X), P(Y)\} = 0.3$. The fact of

Table 5 Comparison of values of *lift*$(X \rightarrow Y)$, *cf*$(X \rightarrow Y)$ and *df*$(X \rightarrow Y)$ when $P(X) + P(Y) \leq 1$

$P(X)$	$P(Y)$	$P(XY)$	$P(X) \times P(Y)$	*lift*$(X \rightarrow Y)$	*cf*$(X \rightarrow Y)$	*cf*$(Y \rightarrow X)$	*df*$(X \rightarrow Y) =$ *df*$(Y \rightarrow X)$
0.60	**0.30**	**0.30**	**0.18**	**1.67**	**0.29**	**1.00**	**1.00**
0.60	0.30	0.25	0.18	1.39	0.17	0.58	0.58
0.60	0.30	0.20	0.18	1.11	0.05	0.17	0.17
0.60	0.30	0.18	0.18	1.00	0.00	0.00	0.00
0.60	0.30	0.15	0.18	0.83	–0.17	–0.17	–0.17
0.60	0.30	0.10	0.18	0.56	–0.44	–0.44	–0.44
0.60	0.30	0.00	0.18	0.00	–1.00	–1.00	–1.00

Table 6 Comparison of values of *lift*$(X \rightarrow Y)$, *cf*$(X \rightarrow Y)$ and *df*$(X \rightarrow Y)$ when $P(X) + P(Y) > 1$

$P(X)$	$P(Y)$	$P(XY)$	$P(X) \times P(Y)$	*lift*$(X \rightarrow Y)$	*cf*$(X \rightarrow Y)$	*cf*$(Y \rightarrow X)$	*df*$(X \rightarrow Y) =$ *df*$(Y \rightarrow X)$
0.80	0.60	0.60	0.48	1.25	0.38	1.00	1.00
0.80	0.60	0.55	0.48	1.15	0.22	0.58	0.58
0.80	0.60	0.50	0.48	1.04	0.06	0.17	0.17
0.80	0.60	0.48	0.48	1.00	0.00	0.00	0.00
0.80	0.60	0.45	0.48	0.94	–0.06	–0.06	–0.37
0.80	**0.60**	**0.40**	**0.48**	**0.83**	**–0.17**	**–0.17**	**–1.00**

reaching the maximal possible value by $P(XY)$ for the given values of $P(X)$ and $P(Y)$ is reflected by the value of *df*$(X \rightarrow Y) = 1$, which means that the dependence between X and Y is strongest possible positive. On the other hand, *cf*$(X \rightarrow Y)$ = 0.29 does not reflect this fact. In general, the real dependence of Y on X may be underestimated when expressed in terms of *cf*$(X \rightarrow Y)$. Also the value 1.67 of *lift*$(X \rightarrow Y)$ itself does not reflect the strong positive dependence between X and Y in the considered case in the view that the lift may reach very large values (close to infinity) in general.

Table 6 shows values of *lift*$(X \rightarrow Y)$, *cf*$(X \rightarrow Y)$ and *df*$(X \rightarrow Y)$ for $P(X) = 0.8$ and $P(Y) = 0.6$; that is, in the case when $P(X) + P(Y) > 1$. For these values of $P(X)$ and $P(Y)$, the minimal possible value of $P(XY)$ equals $P(X) + P(Y) - 1 =$ 0.4. Then the dependence between X and Y is strongest possible negative. This is reflected by the value of *df*$(X \rightarrow Y) = -1$. On the other hand, *cf*$(X \rightarrow Y)$ = –0.17 does not reflect this fact by itself. Also the value 0.83 of *lift*$(X \rightarrow Y)$ itself does not reflect the strong negative dependence between X and Y as it is positioned closer to the value 1 characteristic for independence rather than to the value 0.

5 Dependence Factor as an Interestingness Measure

In [7], Piatetsky-Shapiro postulated that a good interestingness measure of an association rules $X \rightarrow Y$ should fulfill the following conditions:

1. be equal to 0 if X and Y are independent; that is, if $P(XY) = P(X) \times P(Y)$,
2. be increasing with respect to $P(XY)$ given $P(X)$ and $P(Y)$ are fixed,
3. be decreasing with respect to $P(X)$ given $P(XY)$ and $P(Y)$ are fixed or be decreasing with respect to $P(Y)$ given $P(XY)$ and $P(X)$ are fixed.

According to [7], the following rule interest measure $ri(X \rightarrow Y) = |\mathscr{D}| \times [P(XY) - P(X) \times P(Y)]$ fulfills the above postulates. Nevertheless, we notice that this measure does not always satisfy the third postulate. Beneath we present the case in which the *ri* measure violates this postulate:

Let $P(Y) = 0$. Then, $P(XY) = 0$. In this case, $ri(X \rightarrow Y) = 0$ for each value of $P(X)$ in the interval [0, 1]. Thus, $ri(X \rightarrow Y)$ is not guaranteed to be decreasing with respect to $P(X)$ given $P(XY)$ and $P(Y)$ are fixed. Analogically, we would derive that $ri(X \rightarrow Y) = 0$ for each value of $P(Y)$ in the interval [0, 1] if $P(X) = 0$. So, $ri(X \rightarrow Y)$ is not guaranteed to be decreasing with respect to $P(X)$ given $P(XY)$ and $P(Y)$ are fixed. As a result, $ri(X \rightarrow Y)$ does not fulfill the third postulate if $P(X)$ or $P(Y)$ equals 0.

In fact, the $novelty(X \rightarrow Y)$ measure, which was defined in [5] as $[P(XY)–P(X) \times P(Y)]$, violates the third postulate in the same way as $ri(X \rightarrow Y)$.

Now, we will focus on examining if the dependence factor fulfills the postulates of rule interestingness. We start with formulating the properties of probabilities of events which will be useful in our examination.

Proposition 4

(a) *If* $P(X) = 0$ *or* $P(Y) = 0$ *or* $P(X) = 1$ *or* $P(Y) = 1$, *then* $P(XY) = P(X) \times P(Y)$.
(b) *If* $P(XY) \neq P(X) \times P(Y)$, *then* $P(X), P(Y) \in (0, 1)$.

Proof Ad (a) Trivial.

Ad (b) Follows from Proposition 4a. □

Theorem 5 *Let* $X \rightarrow Y$ *be an association rule.*

(a) $df(X \rightarrow Y) = 0$ *iff* $P(XY) = P(X) \times P(Y)$.
(b) *df is increasing with respect to* $P(XY)$ *given* $P(X)$ *and* $P(Y)$ *are fixed.*
(c) *df is non-increasing with respect to* $P(X)$ *given* $P(XY)$ *and* $P(Y)$ *are fixed. In addition, df is decreasing with respect to* $P(X)$ *given* $P(XY)$ *and* $P(Y)$ *are fixed,* $P(Y) \notin \{0, P(XY), 1\}$ *and* $P(XY) \neq 0$.
(d) *df is non-increasing with respect to* $P(Y)$ *given* $P(XY)$ *and* $P(X)$ *are fixed. In addition, df is decreasing with respect to* $P(Y)$ *given* $P(XY)$ *and* $P(X)$ *are fixed,* $P(X) \notin \{0, P(XY), 1\}$ *and* $P(XY) \neq 0$.

Proof Ad (a, b) Follow trivially from Theorems 1 and 3.

Ad (c) Let us first determine the derivative $df'(X \rightarrow Y)$ of $df(X \rightarrow Y)$ as a function of variable $P(X)$ based on Theorem 3 in all possible cases when $P(XY) \neq P(X) \times P(Y)$. We will use the fact that in such cases $P(X), P(Y) \in (0, 1)$ (by Proposition 4b).

Case $P(XY) > P(X) \times P(Y)$ and $\min\{P(X), P(Y)\} = P(X)$.
Then $P(XY) > P(X) \times P(Y) > 0$ and
$df'(X \rightarrow Y) = \frac{P(XY)\times(1-P(Y))}{(P(X)-P(X)\times P(Y))^2} < 0$.

Case $P(XY) > P(X) \times P(Y)$ and $\min\{P(X), P(Y)\} = P(Y)$.
Then
$df'(X \rightarrow Y) = \frac{P(Y)\times(P(XY)-P(Y))}{(P(Y)-P(X)\times P(Y))^2}$.
Hence:

- If $P(XY) = P(Y)$, then $df'(X \rightarrow Y) = 0$.
- If $P(XY) \neq P(Y)$, then $P(XY) < P(Y)$, so $df'(X \rightarrow Y) < 0$.

Case $P(XY) < P(X) \times P(Y)$ and $\max\{0, P(X) + P(Y) - 1\} = 0$.
Then
$df'(X \rightarrow Y) = \frac{PXY\times(P(Y)}{(P(X)\times P(Y))^2}$.
Hence:

- If $P(XY) = 0$, then $df'(X \rightarrow Y) = 0$.
- If $P(XY) \neq 0$, then $df'(X \rightarrow Y) < 0$.

Case $P(XY) < P(X) \times P(Y)$ and $\max\{0, P(X) + P(Y) - 1\} = P(X) + P(Y) - 1$.
Then
$df'(X \rightarrow Y) = \frac{(1-P(Y))\times(P(XY)-P(Y))}{(P(X)\times P(Y)-(P(X)+P(Y)-1))^2} = \frac{(1-P(Y))\times(P(XY)-P(Y))}{((1-P(X))\times(1-P(Y)))^2}$.
Hence:

- If $P(XY) = P(Y)$, then $df'(X \rightarrow Y) = 0$.
- If $P(XY) \neq P(Y)$, then $P(XY) < P(Y)$, so $df'(X \rightarrow Y) < 0$.

Now, let us consider the case when $P(XY) = P(X) \times P(Y)$ and $P(Y) \in (0, 1)$. Then $P(X)$ may take only one value, namely $\frac{P(XY)}{P(Y)}$.

Finally, we note that for $P(Y) = 0$ as well as for $P(Y) = 1$, $P(XY) = P(X) \times P(Y)$ (by Proposition 4a), and so, $df(X \rightarrow Y) = 0$ for each value of $P(X)$ in the interval $[0, 1]$.

Thus, *df* is a non-increasing function with respect to $P(X)$ given $P(XY)$ and $P(Y)$ are fixed. However, if $P(Y) \notin \{0, P(XY), 1\}$ and $P(XY) \neq 0$, then *df* is a decreasing function with respect to $P(X)$ given $P(XY)$ and $P(Y)$ are fixed.
Ad (d) Analogous to the proof of Theorem 5c. □

Corollary 1 *$df(X \rightarrow Y)$ fulfills the first and second Piatetsky-Shapiro postulates. In addition, it fulfills the third Piatetsky-Shapiro postulate if $P(Y) \notin \{0, P(XY), 1\}$ and $P(XY) \neq 0$* or if $P(X) \notin \{0, P(XY), 1\}$ and $P(XY) \neq 0$.

Proof By Theorem 5. □

6 Dependence Factors for Events and Their Complements

In this section, we examine the relationship between the dependence factors for events and their complements. We start with determining extreme values of joint probabilities of events and their complements. Next, we prove that the character of the (in)dependence between X and Y determines uniquely the character of the (in)dependence between $\bar{X}$ and Y, X and $\bar{Y}$, as well as $\bar{X}$ and $\bar{Y}$, respectively. Eventually, we derive the relationship among the dependence factor for X and Y, with those for $\bar{X}$ and Y, X and $\bar{Y}$, as well as $\bar{X}$ and $\bar{Y}$, respectively.

Proposition 5

(a) $max\_P(XY|_{P(X),P(Y)}) = 1$ *iff* $P(X) = P(Y) = 1$.
(b) $min\_P(XY|_{P(X),P(Y)}) = 0$ *iff* $P(X) + P(Y) \leq 1$.
(c) $P(X) + P(Y) \leq 1$ *iff* $(1 - P(X)) + (1 - P(Y)) \geq 1$ *iff* $P(\bar{X}) + P(\bar{Y}) \geq 1$.

Proof Ad (a) Follows from Proposition 1a.
Ad (b) Follows from Proposition 1b.
Ad (c) Trivial. □

Proposition 6

(a) $max\_P(\bar{X}\bar{Y}|_{P(\bar{X}),P(\bar{Y})}) = \min\{P(\bar{X}), P(\bar{Y})\} = \min\{1 - P(X), 1 - P(Y)\} = 1 - \max\{P(X), P(Y)\}$
(b) $min\_P(\bar{X}\bar{Y}|_{P(\bar{X}),P(\bar{Y})}) = \max\{0, P(\bar{X}) + P(\bar{Y}) - 1\} = \max\{0, (1 - P(X)) + (1 - P(Y)) - 1\} = \max\{0, 1 - P(X) - P(Y)\}$
(c) $max\_P(X\bar{Y}|_{P(X),P(\bar{Y})}) = \min\{P(X), P(\bar{Y})\} = min\{P(X), 1 - P(Y)\}$
(d) $min\_P(X\bar{Y}|_{P(X),P(\bar{Y})}) = \max\{0, P(X) + P(\bar{Y}) - 1\} = \max\{0, P(X) + (1 - P(Y)) - 1\} = \max\{0, P(X) - P(Y)\}$
(e) $max\_P(\bar{X}Y|_{P(\bar{X}),P(Y)}) = \min\{P(\bar{X}), P(Y)\} = \min\{1 - P(X), P(Y)\}$
(f) $min\_P(\bar{X}Y|_{P(\bar{X}),P(Y)}) = \max\{0, P(\bar{X}) + P(Y) - 1\} = \max\{0, (1 - P(X)) + P(Y) - 1\} = \max\{0, P(Y) - P(X)\}$

Proof Ad (a, c, e) Follows from Proposition 1a, saying that $max\_P(VZ|_{P(V),P(Z)}) = \min\{P(V), P(Z)\}$.
Ad (b, d, f) Follows Proposition 1b, saying that $min\_P(VZ|_{P(V),P(Z)}) = \max\{0, P(V) + P(Z) - 1\}$. □

Lemma 1

(a) $P(XY) > P(X) \times P(Y)$ *iff* $P(\bar{X}\bar{Y}) > P(\bar{X}) \times P(\bar{Y})$ *iff* $P(X\bar{Y}) < P(X) \times P(\bar{Y})$ *iff* $P(\bar{X}Y) < P(\bar{X}) \times P(Y)$.
(b) $P(XY) = P(X) \times P(Y)$ *iff* $P(\bar{X}\bar{Y}) = P(\bar{X}) \times P(\bar{Y})$ *iff* $P(X\bar{Y}) = P(X) \times P(\bar{Y})$ *iff* $P(\bar{X}Y) = P(\bar{X}) \times P(Y)$.
(c) $P(XY) < P(X) \times P(Y)$ *iff* $P(\bar{X}\bar{Y}) < P(\bar{X}) \times P(\bar{Y})$ *iff* $P(X\bar{Y}) > P(X) \times P(\bar{Y})$ *iff* $P(\bar{X}Y) > P(\bar{X}) \times P(Y)$.

Proof We will proof the proposition using the following equations:

- $P(\bar{X}) = 1 - P(X), \quad P(\bar{Y}) = 1 - P(Y)$,
- $P(\bar{X}Y) = P(Y) - P(XY)$, $P(X\bar{Y}) = P(X) - P(XY)$,
- $P(\bar{X}\bar{Y}) = P(\bar{X}) - P(\bar{X}Y) = 1 - P(X) - P(Y) + P(XY)$.

Ad (a)

- $P(\bar{X}\bar{Y}) > P(\bar{X}) \times P(\bar{Y})$ iff $1 - P(X) - P(Y) + P(XY) > (1 - P(X)) \times (1 - P(Y))$ iff $P(XY) > P(X) \times P(Y)$.
- $P(X\bar{Y}) < P(X) \times P(\bar{Y})$ iff $P(X) - P(XY) < P(X) \times (1 - P(Y))$ iff $P(XY) > P(X) \times P(Y)$.
- $P(\bar{X}Y) < P(\bar{X}) \times P(Y)$ iff $P(Y) - P(XY) < (1 - P(X)) \times P(Y)$ iff $P(XY) > P(X) \times P(Y)$.

Ad (b, c) Analogous to the proof of Lemma 1a. □

Proposition 7

(a) X and Y are dependent positively iff $\bar{X}$ and $\bar{Y}$ are dependent positively iff X and $\bar{Y}$ are dependent negatively iff $\bar{X}$ and Y are dependent negatively.
(b) X and Y are independent iff $\bar{X}$ and $\bar{Y}$ are independent iff X and $\bar{Y}$ are independent iff $\bar{X}$ and Y are independent.
(c) X and Y are dependent negatively iff $\bar{X}$ and $\bar{Y}$ are dependent negatively iff X and $\bar{Y}$ are dependent positively iff $\bar{X}$ and Y are dependent positively.

Proof Follows from Lemma 1. □

Lemma 2 (Proof in Appendix)

(a) $df(X \to Y) = df(\bar{X} \to \bar{Y})$
(b) $df(X \to \bar{Y}) = df(\bar{X} \to Y)$
(c) $df(X \to \bar{Y}) = -df(X \to Y)$

Theorem 6 follows immediately from Lemma 2.

Theorem 6

$$df(X \to Y) = df(\bar{X} \to \bar{Y}) = -df(X \to \bar{Y}) = -df(\bar{X} \to Y).$$

Corollary 2

(a) $df(X \to Y)$ reaches maximum iff $df(\bar{X} \to \bar{Y})$ reaches maximum iff $df(X \to \bar{Y})$ reaches minimum iff $df(\bar{X} \to Y)$ reaches minimum.
(b) $df(X \to Y)$ reaches minimum iff $df(\bar{X} \to \bar{Y})$ reaches minimum iff $df(X \to \bar{Y})$ reaches maximum iff $df(\bar{X} \to Y)$ reaches maximum.

7 Conclusions

In [4], we have offered the dependence factor as a new measure for evaluating the strength of dependence between rules' constituents. Unlike in the case of the certainty factor, when defining the dependence factor, we took into account the fact that for a given rule $X \to Y$, the minimal conditional probability of the occurrence of Y given X may be greater than 0, while its maximal possible value may less than 1. $df(X \to Y)$ always reaches 1 when the dependence between X and Y is strongest possible positive, –1 when the dependence between X and Y is strongest possible negative, and 0 if X and Y are independent. Unlike the dependence factor, the certainty factor itself as well as lift are misleading in expressing the strength of the dependence. In particular, if there is strongest possible positive dependence between X and Y, $cf(X \to Y)$ is not guaranteed to reach its global maximum value 1 (in fact, its value can be quite close to 0 that suggests independence). On the other hand, if there is strongest possible negative dependence between X and Y, $cf(X \to Y)$ is not guaranteed to reach its global minimum value –1 (in fact, its value can be quite close to 0). Similarly, lift may reach values close to the value 1 (that means independence in terms of this measure) even in the cases when the dependence between X and Y is strongest possible positive or strongest possible negative. Thus, we find the dependence factor more accurate measure of a rule constituents' dependence than the certainty factor and lift.

In this paper, we have: (1) examined the dependence factor as an interestingness measure with respect to the interestingness postulates formulated by Piatetsky-Shapiro in [7], and (2) derived the relationship among the dependence factor for X and Y with those for $\bar{X}$ and Y, X and $\bar{Y}$, as well as $\bar{X}$ and $\bar{Y}$, respectively. We have proved that the dependence factor $df(X \to Y)$ fulfills all Piatetsky-Shapiro interestingness postulates if $P(Y) \notin \{0, P(XY), 1\}$ and $P(XY) \neq 0$ or if $P(X) \notin \{0, P(XY), 1\}$ and $P(XY) \neq 0$. Otherwise, it fulfills the first two postulates entirely and the third postulate partially as $df(X \to Y)$ is a non-increasing function rather than decreasing with respect to the marginal probability of an event given the joint probability and the marginal probability of the other event are fixed. On the other hand, it can be observed that several interestingness measures of association rules proposed and/or discussed in the literature does not fulfill all interestingness postulates from [7], including the rule interest *ri* [7] and *novelty* [5], which violate the third postulate for zero marginal probabilities.

In this paper, we have found that the character of the (in)dependence between X and Y determines uniquely the character (positive/negative) of the (in)dependence between $\bar{X}$ and Y, X and $\bar{Y}$, as well as $\bar{X}$ and $\bar{Y}$, respectively. We have also found that the absolute value of the dependence factors is the same for events and their complements. We find this result justified as the marginal and joint probabilities of events and all their complements depend uniquely on the triple of the probabilities $\langle P(X), P(Y), P(XY)\rangle$.

Acknowledgments We wish to thank an anonymous reviewer for constructive comments, which influenced the final version of this paper positively.

Appendix

Proof of Lemma 2
In the proof, we will use the following equations:

- $P(\bar{X}) = 1 - P(X), \qquad P(\bar{Y}) = 1 - P(Y),$
- $P(\bar{X}Y) = P(Y) - P(XY), \quad P(X\bar{Y}) = P(X) - P(XY),$
- $P(\bar{X}\bar{Y}) = P(\bar{X}) - P(\bar{X}Y) = 1 - P(X) - P(Y) + P(XY).$

Ad (a)

Case $P(\bar{X}\bar{Y}) > P(\bar{X}) \times P(\bar{Y})$:

This case is equivalent to the case when $P(XY) > P(X) \times P(Y)$ (by Lemma 1a). Then:

$df(\bar{X} \rightarrow \bar{Y}) =$ /* by Proposition 3a */

$$= \frac{P(\bar{X}\bar{Y}) - P(\bar{X}) \times P(\bar{Y})}{max\_P(\bar{X}\bar{Y}|_{P(\bar{X}),P(\bar{Y})}) - P(\bar{X}) \times P(\bar{Y})} = \text{/* by Proposition 6a */}$$

$$= \frac{(1 - P(X) - P(Y) + P(XY)) - (1 - P(X)) \times (1 - P(Y))}{(1 - \max\{P(X), P(Y)\}) - (1 - P(X)) \times (1 - P(Y))}$$

$$= \frac{P(XY) - P(X) \times P(Y)}{\min\{P(X), P(Y)\} - P(X) \times P(Y)} = \text{/* by Theorem 3 */}$$

$$= df(X \rightarrow Y).$$

Case $P(\bar{X}\bar{Y}) = P(\bar{X}) \times P(\bar{Y})$:

This case is equivalent to the case when $P(XY) = P(X) \times P(Y)$ (by Lemma 1b). Then:

$df(\bar{X} \rightarrow \bar{Y}) =$ /* by Proposition 3a */

$= 0 =$ /* by Proposition 3a */

$= df(X \rightarrow Y).$

Case $P(\bar{X}\bar{Y}) < P(\bar{X}) \times P(\bar{Y})$ and $P(\bar{X}) + P(\bar{Y}) \leq 1$:

This case is equivalent to the case when $P(XY) < P(X) \times P(Y)$ (by Lemma 1c) and $P(X) + P(Y) \geq 1$ (by Proposition 5c). Then:

$df(\bar{X} \rightarrow \bar{Y}) =$ /* by Proposition 3a */

$$- \frac{P(\bar{X}) \times P(\bar{Y}) - P(\bar{X}\bar{Y})}{P(\bar{X}) \times P(\bar{Y}) - min\_P(\bar{X}\bar{Y}|_{P(\bar{X}),P(\bar{Y})})} = \text{/* by Proposition 6b */}$$

$$= - \frac{(1 - P(X)) \times (1 - P(Y)) - 1(-P(X) - P(Y) + P(XY))}{(1 - P(X)) \times (1 - P(Y)) - \max\{0, 1 - P(X), P(Y)\}}$$

$$= - \frac{P(X) \times P(Y) - P(XY)}{(1 - P(X) - P(Y) + P(X) \times P(Y)) - (0)}$$

$$= - \frac{P(X) \times P(Y) - P(XY)}{(P(X) \times P(Y) - (P(X) + P(Y) - 1)}$$

$$
\begin{aligned}
&= -\frac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - \max\{0, P(X) + P(Y) - 1\}} = \text{/* by Theorem 3 */}\\
&= df(X \rightarrow Y).
\end{aligned}
$$

Case $P(\bar{X}\bar{Y}) < P(\bar{X}) \times P(\bar{Y})$ and $P(\bar{X}) + P(\bar{Y}) > 1$:

This case is equivalent to the case when $P(XY) < P(X) \times P(Y)$ (by Lemma 1c) and $P(X) + P(Y) < 1$ (by Proposition 5c). Then:

$df(\bar{X} \rightarrow \bar{Y}) =$ /* by Proposition 3a */

$$
\begin{aligned}
&= -\frac{P(\bar{X}) \times P(\bar{Y}) - P(\bar{X}\bar{Y})}{P(\bar{X}) \times P(\bar{Y}) - min\_P(\bar{X}\bar{Y}|_{P(\bar{X}),P(\bar{Y})})} \text{ /* by Proposition 6b */}\\
&= -\frac{(1 - P(X)) \times (1 - P(Y)) - (1 - P(X) - P(Y) + P(XY))}{(1 - P(X)) \times (1 - P(Y)) - \max\{0, 1 - P(X), P(Y)\}}\\
&= -\frac{P(X) \times P(Y) - P(XY)}{(1 - P(X) - P(Y) + P(X) \times P(Y)) - (1 - P(X) - P(Y))}\\
&= -\frac{P(X) \times P(Y) - P(XY)}{(P(X) \times P(Y) - 0}\\
&= -\frac{P(X) \times P(Y) - P(XY)}{(P(X) \times P(Y) - \max\{0, P(X) + P(Y) - 1\}} = \text{ /* by Theorem 3 */}\\
&= df(X \rightarrow Y).
\end{aligned}
$$

Ad (b)

The proof is analogous to the proof of Lemma 1a.

Ad (c)

Case $P(X\bar{Y}) > P(X) \times P(\bar{Y})$ and $P(X) \leq P(\bar{Y})$:

This case is equivalent to the case when $P(XY) < P(X) \times P(Y)$ (by Lemma 1c) and $P(X) \leq 1 - P(Y)$. Then:

$df(X \rightarrow \bar{Y}) =$ /* by Proposition 3a */

$$
\begin{aligned}
&= \frac{P(X\bar{Y}) - P(X) \times P(\bar{Y})}{max\_P(X\bar{Y}|_{P(X),P(\bar{Y})}) - P(X) \times P(\bar{Y})} = \text{ /* by Proposition 6c */}\\
&= \frac{(P(X) - P(XY)) - P(X) \times (1 - P(Y))}{\min\{P(X), 1 - P(Y)\} - P(X) \times (1 - P(Y))}\\
&= \frac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - 0}\\
&= \frac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - \max\{0, P(X) + P(Y) - 1\}} = \text{ /* by Theorem 3 */}\\
&= -df(X \rightarrow Y).
\end{aligned}
$$

Case $P(X\bar{Y}) > P(X) \times P(\bar{Y})$ and $P(X) > P(\bar{Y})$.

This case is equivalent to the case when $P(XY) < P(X) \times P(Y)$ (by Lemma 1c) and $P(X) > 1 - P(Y)$. Then:

$df(X \rightarrow \bar{Y}) =$ /* by Proposition 3a */

$$= \frac{P(X\bar{Y}) - P(X) \times P(\bar{Y})}{max\_P(X\bar{Y}|_{P(X),P(\bar{Y})}) - P(X) \times P(\bar{Y})} = \text{/* by Proposition 6c */}$$
$$= \frac{(P(X) - P(XY)) - P(X) \times (1 - P(Y))}{\min\{P(X), 1 - P(Y)\} - P(X) \times (1 - P(Y))}$$
$$= \frac{P(X) \times P(Y) - P(XY)}{(1 - P(Y)) - P(X) \times (1 - P(Y))}$$
$$= \frac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - \max\{0, P(X) + P(Y) - 1\}} = \text{/* by Theorem 3 */}$$
$$= -df(X \rightarrow Y).$$

Case $P(X\bar{Y}) = P(X) \times P(\bar{Y})$:

This case is equivalent to the case when $P(XY) = P(X) \times P(Y)$ (by Lemma 1b). Then:

$df(\bar{X} \rightarrow \bar{Y}) =$ /* by Proposition 3a */

$$= 0 = \text{/* by Proposition 3a */}$$
$$= -df(X \rightarrow Y).$$

Case $P(X\bar{Y}) < P(X) \times P(\bar{Y})$ and $P(X) + P(\bar{Y}) \leq 1$.

This case is equivalent to the case when $P(XY) > P(X) \times P(Y)$ (by Lemma 1a) and $P(X) \leq P(Y)$. Then:

$df(X \rightarrow \bar{Y}) =$ /* by Proposition 3a */

$$= -\frac{P(X) \times P(\bar{Y}) - P(X\bar{Y})}{P(X) \times P(\bar{Y}) - min\_P(X\bar{Y}|_{P(X),P(\bar{Y})})} = \text{/* by Proposition 6d */}$$
$$= -\frac{P(X) \times (1 - P(Y)) - (P(X) - P(XY))}{P(X) \times (1 - P(Y)) - \max\{0, P(X) - P(Y)\}}$$
$$= -\frac{P(XY) - P(X) \times P(Y)}{(P(X) - P(X) \times P(Y)) - (0)}$$
$$= -\frac{P(XY) - P(X) \times P(Y)}{\min\{P(X), P(Y)\} - P(X) \times P(Y)} = \text{/* by Theorem 3 */}$$
$$= -df(X \rightarrow Y).$$

Case $P(X\bar{Y}) < P(X) \times P(\bar{Y})$ and $P(X) + P(\bar{Y}) > 1$.

This case is equivalent to the case when $P(XY) > P(X) \times P(Y)$ (by Lemma 1a) and $P(X) > P(Y)$. Then:

$df(X \rightarrow \bar{Y}) =$ /* by Proposition 3a */

$$= -\frac{P(X) \times P(\bar{Y}) - P(X\bar{Y})}{P(X) \times P(\bar{Y}) - min\_P(X\bar{Y}|_{P(X),P(\bar{Y})})} = \text{/* by Proposition 6d */}$$

$$= -\frac{P(X) \times (1 - P(Y)) - (P(X) - P(XY))}{P(X) \times (1 - P(Y)) - \max\{0, P(X) - P(Y)\}}$$

$$= -\frac{P(XY) - P(X) \times P(Y)}{P(X) \times (1 - P(Y)) - (P(X) - P(Y))}$$

$$= -\frac{P(XY) - P(X) \times P(Y)}{P(Y) - P(X) \times P(Y)}$$

$$= -\frac{P(XY) - P(X) \times P(Y)}{\min\{P(X), P(Y)\} - P(X) \times P(Y)} = \text{/* by Theorem 3 */}$$

$$= -df(X \to Y). \qquad \square$$

References

1. Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: ACM SIGMOD international conference on management of data, pp 207–216
2. Brin S, Motwani R, Ullman JD, Tsur S (1997) Dynamic itemset counting and implication rules for market basket data. In: ACM SIGMOD 1997 international conference on management of data, pp 255–264
3. Hilderman RJ, Hamilton HJ (2001) Evaluation of interestingness measures for ranking discovered knowledge. LNCS 2035:247–259
4. Kryszkiewicz M (2015) Dependence factor for association rules. In: Proceedings of ACIIDS 2015, part II LNAI, vol 9012, pp 135–145, Springer, New York
5. Lavrac N, Flach P, Zupan B (1999) Rule evaluation measures: a unifying view. In: Proceedings of ILP-1999. LNAI, vol 1634, pp 174–185. Springer, New York
6. Lenca P, Meyer P, Vaillant B, Lallich S (2008) On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. In: European journal of operational research, vol 184, pp 610–626. Elsevier, France
7. Piatetsky-Shapiro G (1991) Discovery, analysis, and presentation of strong rules. Knowledge discovery in databases, pp 229–248. AAAI/MIT Press, Cambridge
8. Sheikh LM, Tanveer B, Hamdani SMA (2004) Interesting measures for mining association rules. In: Proceedings of INMIC 2004, IEEE
9. Shortliffe E, Buchanan B (1975) A model of inexact reasoning in medicine. Math Biosci 23:351–379
10. Silberschatz A, Tuzhilin A (1995) On subjective measures of interestingness in knowledge discovery. Proc KDD 1995:275–281
11. Suzuki E (2008) Pitfalls for categorizations of objective interestingness measures for rule discovery. In: Statistical implicative analysis: theory and applications, pp 383–395. Springer, New York
12. Suzuki E (2009) Interestingness measures—limits, desiderata, and recent results. In: QIMIE/PAKDD

Recent Results on Nonparametric Quantile Estimation in a Simulation Model

Adam Krzyżak

Abstract We present recent results on nonparametric estimation of a quantile of distribution of Y given by a simulation model $Y = m(X)$, where $m : \mathbb{R}^d \to \mathbb{R}$ is a function which is costly to compute and X is a $\mathbb{R}^d$-valued random variable with given density. We argue that importance sampling quantile estimate of $m(X)$, based on a suitable estimate m_n of m achieves better rate of convergence than the estimate based on order statistics alone. Similar results are given for Robbins-Monro type recursive importance sampling and for quantile estimation based on surrogate model.

1 Introduction

In this paper we consider simulation model of a complex system described by

$$Y = m(X),$$

where X is a $\mathbb{R}^d$-valued random variable with density $f : \mathbb{R}^d \to \mathbb{R}$ and $m : \mathbb{R}^d \to \mathbb{R}$ is an unknown function whose values may be computed at arbitrarily chosen design points, incurring however high computation costs. Let

$$G(y) = \mathbf{P}\{Y \le y\} = \mathbf{P}\{m(X) \le y\} \tag{1}$$

be the cumulative distribution function (cdf) of Y. For $\alpha \in (0, 1)$ we are interested in estimating quantiles of the form

$$q_\alpha = \inf\{y \in \mathbb{R} : G(y) \ge \alpha\}$$

A. Krzyżak (✉)
Department of Computer Science and Software Engineering, Concordia University,
1455 De Maisonneuve Blvd. West, Montreal, QC H3G 1M8, Canada
e-mail: krzyzak@cs.concordia.ca

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_13

using at most n evaluations of the function m. Here we assume that the density f of X is known.

A simple idea is to estimate q_α using an i.i.d. sample $X_1, \ldots, X_n$ of X and to compute the empirical cdf

$$G_{m(X),n}(y) = \frac{1}{n}\sum_{i=1}^{n} I_{\{m(X_i)\leq y\}} \tag{2}$$

and to use the corresponding plug-in estimate

$$\overline{q}_{\alpha,n} = \inf\{y \in \mathbb{R} : G_{m(X),n}(y) \geq \alpha\}. \tag{3}$$

Set $Y_i = m(X_i)$ ($i = 1, \ldots, n$) and let $Y_{1:n}, \ldots, Y_{n:n}$ be the order statistics of $Y_1, \ldots, Y_n$, i.e., $Y_{1:n}, \ldots, Y_{n:n}$ is a permutation of $Y_1, \ldots, Y_n$ such that

$$Y_{1:n} \leq \cdots \leq Y_{n:n}.$$

Since

$$\overline{q}_{\alpha,n} = Y_{\lceil n\alpha\rceil:n}$$

is in fact an order statistic, the properties of this estimate can be studied using the results from order statistics. In particular Theorem 8.5.1 in Arnold et al. [1] implies that in case that $m(X)$ has a density g which is continuous and positive at q_α we have

$$\sqrt{n}\cdot g(q_\alpha)\cdot\frac{Y_{\lceil n\alpha\rceil:n} - q_\alpha}{\sqrt{\alpha\cdot(1-\alpha)}} \to N(0,1) \quad \text{in distribution.}$$

This implies

$$\mathbf{P}\left\{|\bar{q}_{\alpha,n} - q_\alpha| > \frac{c_n}{\sqrt{n}}\right\} \to 0 \quad (n \to \infty) \tag{4}$$

whenever $c_n \to \infty$ ($n \to \infty$).

In this paper we present a survey of our recent results on application of nonparametric techniques to estimating q_α, which lead to faster convergence rates than (4). In particular we will discuss

- nonparametric quantile estimation using importance sampling
- recursive quantile estimation using Robbins-Monro type importance sampling
- nonparametric quantile estimation based on surrogate model.

Throughout this paper we use the following notation: $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{Z}$ and $\mathbb{R}$ are the sets of positive integers, nonnegative integers, integers and real numbers, respectively. For a real number z we denote by $\lfloor z\rfloor$ and $\lceil z\rceil$ the largest integer less than or equal to z and the smallest integer larger than or equal to z, respectively. $\|x\|$ is the Euclidean norm of $x \in \mathbb{R}^d$, and the diameter of a set $A \subseteq \mathbb{R}^d$ is denoted by

$$diam(A) = \sup \{ \|x - z\| \, : \, x, z \in A \} \, .$$

For $f : \mathbb{R}^d \to \mathbb{R}$ and $A \subseteq \mathbb{R}^d$ we set

$$\|f\|_{\infty,A} = \sup_{x \in A} |f(x)|.$$

Let $p = k + s$ for some $k \in \mathbb{N}_0$ and $0 < s \leq 1$, and let $C > 0$. A function $m : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-smooth, if for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = k$ the partial derivative $\frac{\partial^k m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ exists and satisfies

$$\left| \frac{\partial^k m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) - \frac{\partial^k m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|^s$$

for all $x, z \in \mathbb{R}^d$.

For nonnegative random variables X_n and Y_n we say that $X_n = O_{\mathbf{P}}(Y_n)$ if

$$\lim_{c \to \infty} \limsup_{n \to \infty} \mathbf{P}(X_n > c \cdot Y_n) = 0.$$

The paper is organized as follows. In Sect. 2 we introduce importance sampling and apply it to quantile estimation. Quantile estimation by recursive procedure of Robbins-Monro type importance sampling is discussed in Sect. 3 and Monte Carlo surrogate quantile estimates are considered in Sect. 4.

2 Nonparametric Quantile Estimation Using Importance Sampling

In this section we apply importance sampling (IS) to obtain an estimate of q_α which converges faster to q_α than order statistics. Our presentation is based on Kohler et al. [31] where more detailed treatment of the problem at hand is given. Given a sequence of independent and identically distributed random variables $X, X_1, X_2, \ldots$ and a function $\phi : \mathbb{R}^d \to \mathbb{R}$ the standard approach to estimating $\mathbf{E}\phi(X)$ is to use sample averages

$$\frac{1}{n} \sum_{i=1}^n \phi(X_i).$$

The goal of importance sampling is to improve estimation of $\mathbf{E}\phi(X)$, by using a new random variable Z with a density h satisfying for all $x \in \mathbb{R}^d$

$$\phi(x) \cdot f(x) \neq 0 \quad \Rightarrow \quad h(x) \neq 0$$

and generating an i.i.d. sequence $Z, Z_1, Z_2, \dots$ which is then used to estimate

$$\mathbf{E}\{\phi(X)\} = \mathbf{E}\left\{\phi(Z) \cdot \frac{f(Z)}{h(Z)}\right\}$$

by

$$\frac{1}{n}\sum_{i=1}^{n} \phi(Z_i) \cdot \frac{f(Z_i)}{h(Z_i)}, \tag{5}$$

whereas we assume that $\frac{0}{0} = 0$. We choose h such that the variance of (5) is small (see for instance Sect. 4.6 in Glasserman [18], Neddermayer [38] and the literature cited therein).

Quantile estimation using importance sampling has been considered by Cannamela et al. [6], Egloff and Leippold [16] and Morio [35]. The authors proposed new estimates in various models, however only Egloff and Leippold [16] investigated consistency of their method. None of the papers analyzed the rates of convergence.

As m is costly to evaluate we replace it by a surrogate function which is cheap to evaluate at arbitrary points and we use important sampling to sample it using of cluster properties of the distribution of X. Such surrogate functions have been considered in context of quadratic response surfaces by Kim and Na [25] and Das and Zheng [8], in context of support vector machines by Hurtado [23], Deheeger and Lemaire [10] and Bourinet et al. [5], in context of neural networks by Papadrakakis and Lagaros [40], and in context of kriging and related concepts by Santner et al. [45] (see Sects. 3.3 and 3.4 with further references therein), Kaymaz [24] and Bichon et al. [4]. No theoretical results concerning the rates of convergence of the investigated estimates have been provided in the papers cited above.

Oakley [39] and Dubourg et al. [15] use the kriging approximation method, with pragmatic devices, the latter authors for the related, but simpler problem of estimating the distribution function value of $m(X)$ at 0 (i.e., estimation of a failure rate, here 0 instead of a general real y without loss of generality). They assume parametric model with m that is a realization of a Gaussian distributed random function. On the basis of a set of pairs $(x_1, y_1), \dots, (x_n, y_n)$ of moderate size n, with design vectors x_i suitably chosen and $y_i = m(x_i)$, by the Bayes principle a posterior distribution for m is obtained, with posterior mean function m^*, which interpolates m at the design points $x_1, \dots, x_n$. According to this posterior distribution a new realization function is chosen with cheap evaluations on a large set of randomly chosen design vectors. This yields more information on the original estimation problem including some useful additional design points and the procedure is repeated. Refer to [15, 39] for more details.

In Kohler et al. [31] a new importance sampling quantile estimate is proposed and its rates of convergence are analyzed. It is done in a completely nonparametric context imposing mild smoothness assumption on m (the structure of m is unknown) in view of good approximation by a surrogate function. The basic idea is to use an initial estimate of the quantile based on the order statistics of samples of $m(X)$ in order to determine an interval $[a_n, b_n]$ containing the quantile. Then we construct an estimate m_n of m and restrict f to the inverse image $m_n^{-1}([a_n, b_n])$ of $[a_n, b_n]$ to construct a new random variable Z enabling us to sample only from the region important for the computation of the quantile. Our final estimate of the quantile is then defined as an order statistic of $m(Z)$, where the level of the order statistics takes into account that we sample only from a part of the original density f. Under suitable assumptions on the smoothness of m and on the tails of f we are able to show that this estimate achieves the rate of convergence of order $\frac{\log^{1.5} n}{n}$.

2.1 Definition of the Estimate

Let $n = n_1 + n_2 + n_3$ where $n_1 = n_1(n) = \lfloor n/3 \rfloor = n_2 = n_2(n)$ and $n_3 = n_3(n) = n - n_1 - n_2$. We use n_1 evaluations of m to generate an initial estimate of q_α, n_2 evaluations of m to construct an approximation of m, and we use n_3 additional evaluations of m to improve our initial estimate of q_α.

Let $\overline{q}_{\alpha,n_1}$ be the quantile estimate based on order statistics introduced in Sect. 1. In order to improve it by importance sampling, we will use additional observations $(x_1, m(x_1))$, ..., $(x_{n_2}, m(x_{n_2}))$ of m at points $x_1, \dots, x_{n_2} \in \mathbb{R}^d$ and use an estimate

$$m_n(\cdot) = m_n(\cdot, (x_1, m(x_1)), \dots, (x_{n_2}, m(x_{n_2}))) : \mathbb{R}^d \to \mathbb{R}$$

of $m : \mathbb{R}^d \to \mathbb{R}$. Both will be specified later. Let $K_n = [-l_n, l_n]^d$ for some $l_n > 0$ such that $l_n \to \infty$ as $n \to \infty$ and assume that the supremum norm error of m_n on K_n is bounded by $\beta_n > 0$, i.e.,

$$\|m_n - m\|_{\infty,K_n} := \sup_{x \in K_n} |m_n(x) - m(x)| \le \beta_n. \tag{6}$$

Set

$$a_n = \overline{q}_{\alpha,n_1} - 2 \cdot \frac{\log n}{\sqrt{n}} - 2 \cdot \beta_n \quad \text{and} \quad b_n = \overline{q}_{\alpha,n_1} + 2 \cdot \frac{\log n}{\sqrt{n}} + \beta_n,$$

where both quantities depend (via $\overline{q}_{\alpha,n_1}$) on the data

$$\mathrm{d}_{n_1} = \left\{(X_1, m(X_1)), \dots, (X_{n_1}, m(X_{n_1}))\right\}.$$

Next we replace X by a random variable Z with the density

$$h(x) = c_2 \cdot \left(I_{\{x \in K_n : a_n \le m_n(x) \le b_n\}} + I_{\{x \notin K_n\}}\right) \cdot f(x)$$

where

$$c_2 = \left(\int_{\mathbb{R}^d} \left(I_{\{x \in K_n : a_n \le m_n(x) \le b_n\}} + I_{\{x \notin K_n\}} \right) f(x) dx \right)^{-1} = \frac{1}{1 - \gamma_1 - \gamma_2}.$$

Here

$$\gamma_1 = \mathbf{P}\{X \in K_n, m_n(X) < a_n | \mathrm{d}_{n_1}\} = \int_{\mathbb{R}^\mathrm{d}} 1_{\mathrm{K_n}}(\mathrm{x}) \cdot 1_{\{\mathrm{x}\,:\,\mathrm{m_n}(\mathrm{x}) < \mathrm{a_n}\}} \cdot \mathrm{f}(\mathrm{x})\mathrm{dx}$$

and

$$\gamma_2 = \mathbf{P}\{X \in K_n, m_n(X) > b_n | \mathrm{d}_{n_1}\} = \int_{\mathbb{R}^\mathrm{d}} 1_{\mathrm{K_n}}(\mathrm{x}) \cdot 1_{\{\mathrm{x}\,:\,\mathrm{m_n}(\mathrm{x}) > \mathrm{b_n}\}} \cdot \mathrm{f}(\mathrm{x})\mathrm{dx}$$

can be computed exactly for given f and m_n. A realization of random variable Z can be constructed by a rejection method: We generate independent realizations of X until we observe a realization x which satisfies either $x \in [-l_n, l_n]^d$ and $a_n \le m_n(x) \le b_n$ or $x \notin [-l_n, l_n]^d$, which we then use as the realization of our Z_n. In our application below we approximate them by the suitable Riemann sums. Observe that a_n and b_n depend on d_{n_1} and therefore the density h and the distribution of Z are random quantities. Furthermore on the event

$$\left\{ |\overline{q}_{\alpha,n_1} - q_\alpha| \le \frac{\log n}{\sqrt{n}} \right\}$$

we have that

$$\int_{\mathbb{R}^d} \left(I_{\{x \in K_n : a_n \le m_n(x) \le b_n\}} + I_{\{x \notin K_n\}} \right) f(x) dx \ge \mathbf{P} \left\{ q_\alpha - \frac{\log n}{\sqrt{n}} \le m(X) \le q_\alpha + \frac{\log n}{\sqrt{n}} \right\} > 0, \tag{7}$$

provided, e.g., the density of $m(X)$ is positive and continuous at q_α. Hence outside of an event whose probability tends to zero for $n \to \infty$ the constant c_2 and the density h are in this case well defined. A key Lemma 1 below relates the quantile q_α to the quantile of $m(Z)$ (for the proof refer to [31]).

Lemma 1 *Assume that* (6) *holds,* $m(X)$ *has a density which is continuous and positive at* q_α *and let* Z *be a random variable defined as above. Furthermore set*

$$\bar{\alpha} = \frac{\alpha - \gamma_1}{1 - \gamma_1 - \gamma_2}$$

and

$$q_{m(Z),\bar{\alpha}} = \inf\{y \in \mathbb{R} \,:\, \mathbf{P}\{m(Z) \le y | \mathrm{d}_{\mathrm{n_1}}\} \ge \bar{\alpha}\}$$

where $\mathrm{d}_{n_1} = \{(X_1, m(X_1)), \ldots, (X_{n_1}, m(X_{n_1})\}$. *Then we have with probability tending to one for* $n \to \infty$ *that*

$$q_\alpha = q_{m(Z),\bar{\alpha}}.$$

Let Z, Z_1, Z_2, …be independent and identically distributed and set

$$G_{m(Z),n_3}(y) = \frac{1}{n_3} \sum_{i=1}^{n_3} I_{\{m(Z_i) \le y\}}.$$

We estimate q_α (which is outside of an event whose probability tends to zero for $n \to \infty$ according to Lemma 1 equal to $q_{m(Z),\bar{\alpha}}$) by

$$\begin{aligned}\bar{q}_{m(Z),\bar{\alpha},n_3} &= \inf\left\{y \in \mathbb{R} \,:\, G_{m(Z),n_3}(y) \ge \bar{\alpha}\right\} \\ &= \inf\left\{y \in \mathbb{R} \,:\, G_{m(Z),n_3}(y) \ge \frac{\alpha - \gamma_1}{1 - \gamma_1 - \gamma_2}\right\}.\end{aligned}$$

As before we have that $\bar{q}_{m(Z),\bar{\alpha},n_3}$ is an order statistic of $m(Z_1)$, …, $m(Z_{n_3})$:

$$\bar{q}_{m(Z),\bar{\alpha},n_3} = m(Z)_{\lceil \bar{\alpha} \cdot n_3 \rceil : n_3}.$$

we approximate m by the spline estimate m_n introduced below. We use well-known results from spline theory to show that if we choose the design points z_1, …, z_n equidistantly in $K_n = [-l_n, l_n]^d$, then a properly defined spline approximation of a (p, C)-smooth function achieves the rate of convergence $l_n^p / n^{p/d}$.

In order to define the spline approximation, we introduce polynomial splines, i.e., sets of piecewise polynomials satisfying a global smoothness condition, and a corresponding B-spline basis consisting of basis functions with compact support as follows:

Choose $K \in \mathbb{N}$ and $M \in \mathbb{N}_0$, and set $u_k = k \cdot l_n / K$ $(k \in \mathbb{Z})$. For $k \in \mathbb{Z}$ let $B_{k,M} : \mathbb{R} \to \mathbb{R}$ be the univariate B-spline of degree M with knot sequence $(u_k)_{k \in \mathbb{Z}}$ and support $supp(B_{k,M}) = [u_k, u_{k+M+1}]$. In case $M = 0$ B-spline $B_{k,0}$ is the indicator function of the interval $[u_k, u_{k+1})$, and for $M = 1$ we have

$$B_{k,1}(x) = \begin{cases} \frac{x - u_k}{u_{k+1} - u_k} & , u_k \le x \le u_{k+1}, \\ \frac{u_{k+2} - x}{u_{k+2} - u_{k+1}} & , u_{k+1} < x \le u_{k+2}, \\ 0 & , \text{elsewhere}, \end{cases}$$

(so-called hat-function). The general recursive definition of $B_{k,M}$ can be found, e.g., in de Boor [9], or in Sect. 14.1 of Györfi et al. [21]. These B-splines are basis functions of sets of univariate piecewise polynomials of degree M, where the piecewise polynomials are globally $(M - 1)$-times continuously differentiable and where the Mth derivatives of the functions have jump points only at the knots u_l $(l \in \mathbb{Z})$.

For $\mathbf{k}=(k_1,\ldots,k_d)\in\mathbb{Z}^d$ we define the tensor product B-spline $B_{\mathbf{k},M}:\mathbb{R}^d\to\mathbb{R}$ by

$$B_{\mathbf{k},M}(x^{(1)},\ldots,x^{(d)})=B_{k_1,M}(x^{(1)})\cdot\ldots\cdot B_{k_d,M}(x^{(d)})\quad(x^{(1)},\ldots,x^{(d)}\in\mathbb{R}).$$

With these functions we define $\mathscr{S}_{K,M}$ as the set of all linear combinations of all those tensor product B-splines above, whose support has nonempty intersection with $K_n=[-l_n,l_n]^d$, i.e., we set

$$\mathscr{S}_{K,M}=\left\{\sum_{\mathbf{k}\in\{-K-M,-K-M+1,\ldots,K-1\}^d}a_{\mathbf{k}}\cdot B_{\mathbf{k},M}\ :\ a_{\mathbf{k}}\in\mathbb{R}\right\}.$$

It can be shown by using standard arguments from spline theory, that the functions in $\mathscr{S}_{K,M}$ are in each component $(M-1)$-times continuously differentiable and that they are equal to a (multivariate) polynomial of degree less than or equal to M (in each component) on each rectangle

$$[u_{k_1},u_{k_1+1})\times\cdots\times[u_{k_d},u_{k_d+1})\quad(\mathbf{k}=(k_1,\ldots,k_d)\in\mathbb{Z}^d),\tag{8}$$

and that they vanish outside the set

$$\left[-l_n-M\cdot\frac{l_n}{K},l_n+M\cdot\frac{l_n}{K}\right]^d.$$

Next we define spline approximations using so-called quasi interpolants: For a continuous function $m:\mathbb{R}^d\to\mathbb{R}$ we define an approximating spline by

$$(Qm)(x)=\sum_{\mathbf{k}\in\{-K-M,-K-M+1,\ldots,K-1\}^d}Q_{\mathbf{k}}m\cdot B_{\mathbf{k},M}$$

where

$$Q_{\mathbf{k}}m=\sum_{\mathbf{j}\in\{0,1,\ldots,M\}^d}a_{\mathbf{k},\mathbf{j}}\cdot m(t_{k_1,j_1},\ldots,t_{k_d,j_d})$$

for some $a_{\mathbf{k},\mathbf{j}}\in\mathbb{R}$ defined below and some suitably chosen points $t_{k,j}\in supp(B_{k,M})=[k\cdot l_n/K,(k+M+1)\cdot l_n/K]$. It can be shown that if we set

$$t_{k,j}=k\cdot\frac{l_n}{K}+\frac{j}{M}\cdot\frac{l_n}{K}\quad(j\in\{0,\ldots,M\},k\in\{-K,-K+1,\ldots,K-1\})$$

and

$$t_{k,j}=-l_n+\frac{j}{M}\cdot\frac{l_n}{K}\quad(j\in\{0,\ldots,M\},k\in\{-K-M,-K-M+1,\ldots,-K-1\}),$$

then there exist coefficients $a_{\mathbf{k},\mathbf{j}}$ (which can be computed by solving a linear equation system), such that

$$|Q_{\mathbf{k}} f| \leq c_3 \cdot \|f\|_{\infty,[u_{k_1},u_{k_1+M+1}]\times\cdots\times[u_{k_d},u_{k_d+M+1}]} \tag{9}$$

for any $\mathbf{k} \in \mathbb{Z}^d$, any continuous $f : \mathbb{R}^d \to \mathbb{R}$ and some universal constant c_1, and such that Q reproduces polynomials of degree M or less (in each component) on $K_n = [-l_n, l_n]^d$, i.e., for any multivariate polynomial $p : \mathbb{R}^d \to \mathbb{R}$ of degree M or less in each component we have

$$(Qp)(x) = p(x) \quad (x \in K_n) \tag{10}$$

(cf., e.g., Theorems 14.4 and 15.2 in [21]).

Next we define our estimate m_n as a quasi interpolant. We fix the degree $M \in \mathbb{N}$ and set

$$K = \left\lfloor \frac{\lfloor n_2^{1/d} \rfloor - 1}{2M} \right\rfloor,$$

where we assume that $n_2 \geq (2M+1)^d$. Furthermore we choose $x_1, \ldots, x_{n_2}$ such that all of the $(2M \cdot K + 1)^d$ points of the form

$$\left(\frac{j_1}{M \cdot K} \cdot l_n, \ldots, \frac{j_d}{M \cdot K} \cdot l_n \right) \quad (j_1, \ldots, j_d \in \{-M \cdot K, -M \cdot K + 1, \ldots, M \cdot K\})$$

are contained in $\{x_1, \ldots, x_{n_2}\}$, which is possible since $(2M \cdot K + 1)^d \leq n_2$. Then we define

$$m_n(x) = (Qm)(x),$$

where Qm is the above defined quasi interpolant satisfying (17) and (19). The computation of Qm requires only function values of m at the points $x_1, \ldots, x_{n_2}$ and hence m_n is well defined.

It follows from spline theory (cf., e.g., proof of Theorem 1 in Kohler [29]) that if m is (p, C)-smooth for some $0 < p \leq M + 1$ then the above quasi interpolant m_n satisfies for some constant $c_4 > 0$

$$\|m_n - m\|_{\infty,K_n} := \sup_{x \in K_n} |m_n(x) - m(x)| \leq c_4 \cdot \frac{l_n^p}{n_2^{p/d}}, \tag{11}$$

i.e., (6) is satisfied with $\beta_n = c_4 \cdot l_n^p / n_2^{p/d}$.

2.2 Main Results

The following theorem presents the rate of convergence result for the quantile estimate using a general estimate of m.

Theorem 1 *Assume that X is a $\mathbb{R}^d$-valued random variable which has a density with respect to the Lebesgue measure. Let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function. Assume that $m(X)$ has a density g with respect to the Lebesgue measure and let $\alpha \in (0, 1)$. Assume that the density g of $m(X)$ is positive at q_α and continuous on $\mathbb{R}$.*

Let the estimate $\bar{q}_{Z,\bar{\alpha},n}$ of q_α be defined as in Sect. 2.1 with $\beta_n = \frac{\log n}{\sqrt{n}}$ and assume that regression estimate m_n satisfies (6). Furthermore assume that

$$\mathbf{P}\{X \notin K_n\} = O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right) \tag{12}$$

Then

$$|\bar{q}_{m(Z),\bar{\alpha},n_3} - q_\alpha| = O_{\mathbf{P}}\left(\frac{\log^{1.5}(n)}{n}\right).$$

The proof is given in [31] and is omitted. For m estimated by the spline estimate from the previous section we get the following result.

Corollary 1 *Assume that X is a $\mathbb{R}^d$-valued random variable which has a density with respect to the Lebesgue measure. Let $m : \mathbb{R}^d \to \mathbb{R}$ be a (p, C)-smooth function for some $p > d/2$. Assume that $m(X)$ has a density g with respect to the Lebesgue measure. Let $\alpha \in (0, 1)$ and let q_α be the α-quantile of $m(X)$. Assume that the density g of $m(X)$ is positive at q_α and continuous on $\mathbb{R}$.*

Let m_n be the spline estimate from Sect. 2.1 with $M \geq p - 1$ and define the estimate $\bar{q}_{Z,\bar{\alpha},n}$ of q_α as in Sect. 2.1 with $\beta_n = \frac{\log n}{\sqrt{n}}$ and $l_n = \log n$. Furthermore assume that

$$\mathbf{P}\{||X|| \geq \log n\} = O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right). \tag{13}$$

Then

$$|\bar{q}_{m(Z),\bar{\alpha},n_3} - q_\alpha| = O_{\mathbf{P}}\left(\frac{\log^{1.5}(n)}{n}\right).$$

It follows from Markov inequality that (13) is satisfied whenever

$$\mathbf{E}\left\{\exp\left(\frac{1}{2} \cdot \|X\|\right)\right\} < \infty.$$

If (13) does not hold it is possible to change the definition of l_n in Corollary 1 to get an (maybe modified) assertion under a weaker tail condition. It is possible to improve the factor $\log^{1.5}(n)$ in Corollary 1, provided one changes the definition of a_n and b_n. More precisely, let $(\gamma_n)_n$ be a monotonically increasing sequence of positive real values which tends to infinity and assume

$$\mathbf{P}\{||X|| \geq \log n\} = O\left(\frac{\sqrt{\gamma_n}}{\sqrt{n}}\right).$$

Set

$$a_n = \overline{q}_{\alpha,n_1} - \frac{\sqrt{\gamma_n}}{\sqrt{n}} \quad \text{and} \quad b_n = \overline{q}_{\alpha,n_1} + \frac{\sqrt{\gamma_n}}{\sqrt{n}}.$$

By applying (4) in the proof of Theorem 1 it is possible to show that under the assumptions of Corollary 1 the estimate based on the above modified values of a_n and b_n satisfies

$$|\bar{q}_{m(Z),\bar{\alpha},n} - q_\alpha| = O_{\mathbf{P}}\left(\frac{\gamma_n}{n}\right).$$

3 Recursive Quantile Estimation Using Robbins-Monro Type Importance Sampling

Here we summarize our results on recursive quantile estimation using Robbins-Monro importance sampling. For full account we refer the reader to Kohler et al. [30]. In this section we use ideas from Kohler et al. [30] and importance sampling combined with an approximation of the underlying function m in order to improve the rate of convergence of our recursive estimate of the quantile. Let Y be a real-valued random variable with cumulative distribution function (cdf) $G(y) = \mathbf{P}\{Y \leq y\}$. We are interested in estimating quantiles of Y of level $\alpha \in (0, 1)$, which can be defined as any value between

$$q_\alpha^{lower} = \inf\{y \in \mathbb{R} : G(y) \geq \alpha\} \quad \text{and} \quad q_\alpha^{upper} = \sup\{y \in \mathbb{R} : G(y) \leq \alpha\}.$$

We assume that Y has a bounded density g with respect to the Lebesgue-Borel-measure which is positive in a neighborhood of q_α^{upper}, which implies that there exists a uniquely determined quantile $q_\alpha = q_\alpha^{upper} = q_\alpha^{lower}$. Let Y, Y_1, Y_2, …be independent and identically distributed. Given Y_1, …, Y_n, we are interested in estimates $\hat{q}_{n,\alpha} = \hat{q}_{n,\alpha}(Y_1, \ldots, Y_n)$ of q_α with the property that the error $\hat{q}_{n,\alpha} - q_\alpha$ converges quickly towards zero in probability as $n \to \infty$.

A simple estimate of q_α is given by order statistics. Let $Y_{1:n}, \ldots, Y_{n:n}$ be the order statistics of $Y_1, \ldots, Y_n$, i.e., $Y_{1:n}, \ldots, Y_{n:n}$ is a permutation of $Y_1, \ldots, Y_n$ such that $Y_{1:n} \leq \ldots \leq Y_{n:n}$. Then we can estimate q_α by

$$\overline{q}_{\alpha,n} = Y_{\lceil n\alpha \rceil:n}.$$

Note that $\overline{q}_{\alpha,n}$ satisfies (4).

In order to compute the above estimate one needs to sort the given data Y_1, ..., Y_n in increasing order, which requires an amount of time of order $n \cdot \log(n)$ and an amount of space of order n (the latter one in order to save all values of the data points simultaneously). In case that one wants to compute a quantile estimate for a very large sample size, a recursive estimate might be more appropriate. Such a recursive estimate can be computed by applying the Robbins-Monro procedure to estimate the root of $G(z) - \alpha$. In its most simple form one starts here with an arbitrary random variable Z_1, e.g., $Z_1 = 0$, and defines the quantile estimate Z_n recursively via

$$Z_{n+1} = Z_n - \frac{D_n}{n} \cdot \left(I_{\{Y_n \le Z_n\}} - \alpha \right) \tag{14}$$

for some suitable sequence $D_n \ge 0$. The Robbins-Monro procedure was originally proposed by Robbins and Monro [43] and further developed and investigated as well as applied in many different situations, cf., e.g., the monographs Benveniste et al. [3], Ljung et al. [33], Chen [7] and Kushner and Yin [32], and the literature cited therein. Refined versions of the above simple Robbins-Monro estimate achieve the same rate of convergence as in (4) and (4), explicitly in Tierney [49] and Holst [22] by additional use of a recursive estimate of $g(q_\alpha)$ or, for g Hölder continuous at q_α, as a consequence of general results on averaged Robbins-Monro estimates due to Ruppert [44] and Polyak and Juditsky [41].

Consider again simulation model (1), where random variable Y is given by $Y = m(X)$ for some known measurable function $m : \mathbb{R}^d \to \mathbb{R}$ and some $\mathbb{R}^d$-valued random variable X. We want to use an importance sampling variant of the recursive estimate (14) based on a suitably defined approximation m_n of m. In case that the function m is p-times continuously differentiable and that X satisfies a proper exponential moment condition we show that this importance sampling variant of the recursive estimate achieves up to some logarithmic factor a rate of convergence of order $n^{-1/2-p/(2d)}$ for $0 < p \le d$.

3.1 Main Result

We combine a Robbins-Monro estimate with importance sampling in order to improve the rate of convergence. Here we assume that our data is given by $Y = m(X)$ for some known measurable function $m : \mathbb{R}^d \to \mathbb{R}$ and some $\mathbb{R}^d$-valued random variable X with known distribution μ. We assume that we have available a deterministic approximation $\tilde{m}_n$ of m which satisfies

$$\|\tilde{m}_n - m\|_{\infty,[-l_n,l_n]^d} \le \log^{p+1}(n) \cdot n^{-p/d} \tag{15}$$

for sufficiently large n for some $0 < p \le d$, where $l_n = \log(n)$. Set

$$m_n = \tilde{m}_n - \log^{p+1}(n) \cdot n^{-p/d}.$$

Then we have

$$\|m_n - m\|_{\infty,[-l_n,l_n]^d} \leq 2 \cdot \log^{p+1}(n) \cdot n^{-p/d} \tag{16}$$

and

$$m_n(x) \leq m(x) \quad \text{for all } x \in [-l_n, l_n]^d \tag{17}$$

for sufficiently large n (more precisely, for $n \geq n_0$, where $n_0 \in \mathbb{N}$ is some unknown positive deterministic integer).

We recursively define a sequence of estimates Z_n of q_α. We start by choosing an arbitrary (w.l.o.g. deterministic) Z_1, e.g., $Z_1 = 0$. After having constructed already Z_1, ..., Z_n, we choose a random variable $X_n^{(IS)}$ such that $X_n^{(IS)}$ has the distribution

$$H_n(B) = \frac{\mu\left(\left(\{x \in [-l_n, l_n]^d \ : \ m_n(x) \leq Z_n\} \cup ([-l_n, l_n]^d)^c\right) \cap B\right)}{\bar{G}_n(Z_n)} \quad (B \in \mathscr{B}^d)$$

where $\mathscr{B}^d$ is the set of all Borel sets in $\mathbb{R}^d$ and where

$$\bar{G}_n(z) = \mu\left(\{x \in [-l_n, l_n]^d \ : \ m_n(x) \leq z\} \cup ([-l_n, l_n]^d)^c\right). \tag{18}$$

By construction, the distribution H_n has the Radon-Nikodym derivative (conditional on Z_n)

$$\frac{dH_n}{d\mu}(x) = \frac{I_{\{m_n(x) \leq Z_n\}} \cdot I_{\{x \in [-l_n,l_n]^d\}} + I_{\{x \notin [-l_n,l_n]^d\}}}{\bar{G}_n(Z_n)}.$$

A realization of such a random variable can be constructed using a rejection method: we generate independent realizations of X until we observe a realization x which satisfies either $x \in [-l_n, l_n]^d$ and $m_n(x) \leq Z_n$ or $x \notin [-l_n, l_n]^d$, which we then use as the realization of our $X_n^{(IS)}$.

Next we choose i.i.d. random variables $X_{n,1}$, $X_{n,2}$, ..., $X_{n,n}$ distributed as X, which are independent of all other random variables constructed or used until this point and we set

$$Z_{n+1} = Z_n - \frac{D_n}{n} \cdot \left(I_{\{m(X_n^{(IS)}) \leq Z_n\}} \cdot \tilde{G}_n(Z_n) - \alpha\right), \tag{19}$$

where $D_n = \log^2(n)$ and

$$\tilde{G}_n(z) = \frac{1}{n} \sum_{i=1}^{n} \left(I_{\{m_n(X_{n,i}) \leq z, X_{n,i} \in [-l_n,l_n]^d\}} + I_{\{X_{n,i} \notin [-l_n,l_n]^d\}}\right) \quad (z \in \mathbb{R}).$$

The main result below is an upper bound on the error of this quantile estimate.

Theorem 2 *Let X, $X_{1,1}$, $X_{2,1}$, $X_{2,2}$, $X_{3,1}$, $X_{3,2}$, $X_{3,3}$, ...be independent and identically distributed $\mathbb{R}^d$-valued random variables and let $m : \mathbb{R}^d \rightarrow \mathbb{R}$ be a* *measurable function. Let* $\alpha \in (0, 1)$ *and let* q_α *be the* α*-quantile of* $Y = m(X)$. *Assume that* $Y = m(X)$ *has a bounded density* g *with respect to the Lebesgue-Borel measure which is bounded away from zero in a neighborhood of* q_α. *Define* $X_n^{(IS)}$ *as above, where* m_n *satisfies (16) and (17) for some* $0 < p \leq d$, *and let* $\hat{q}_{\alpha,n}^{(IS)} = Z_n$ *be the Robbins-Monro importance sampling quantile estimate defined above with* $D_n = \log^2(n)$. *Then*

$$\mathbf{P}\left\{X \notin [-\log(n), \log(n)]^d\right\} > 0 \quad (n \in \mathbb{N}) \quad \textit{and} \quad \mathbf{E}\{e^{\|X\|}\} < \infty \tag{20}$$

imply

$$\hat{q}_{\alpha,n}^{(IS)} \to q_\alpha \quad a.s. \quad \textit{and} \quad \left|\hat{q}_{\alpha,n}^{(IS)} - q_\alpha\right| = O_{\mathbf{P}}\left(\log^{3+p/2}(n) \cdot n^{-1/2-p/(2d)}\right).$$

The proof can be found in [30] and is omitted.

Remark 1 The construction of an approximation m_n which satisfies (15) in case of a (p, C)—smooth function m can be obtained, e.g., by spline approximation of the function m using n points in $[-\log(n), \log(n)]^d$ (cf., e.g., Kohler et al. [31]), which can be either chosen equidistantly in $[-\log(n), \log(n)]^d$ or can be defined recursively such that we reuse for computation of m_{n+1} evaluations of m used for computation of m_n. Thus as in the case importance sampling algorithm discussed in Kohler et al. [31], our algorithm achieves the faster rate of convergence than the estimate based on order statistics, but it requires less space to be computed than the order statistics or the estimate in Kohler et al. [31].

4 Nonparametric Quantile Estimation Based on Surrogate Model

In this section we present quantile estimates based on a surrogate model. These estimates achieve under suitable conditions better rates of convergence than the ones based on order statistics (see (4)). For in-depth discussion we refer the reader to Enss et al. [17]. The basic idea is to first construct an estimate m_n of m and then to estimate the quantile $q_{m(X),\alpha}$ by a Monte Carlo estimate of the quantile $q_{m_n(X),\alpha}$, where

$$q_{m_n(X),\alpha} = \inf\left\{y \in \mathbb{R} : \mathbf{P}_X\{x \in \mathbb{R}^d : m_n(x) \leq y\} \geq \alpha\right\}.$$

The main result presented in Theorem 3 below concerns the error of this Monte Carlo estimate. We show that if the local error of m_n is small in areas where $m(x)$ is close to $q_{m(X),\alpha}$, i.e., if for some small $\delta_n > 0$

$$|m_n(x) - m(x)| \leq \frac{\delta_n}{2} + \frac{1}{2} \cdot |m(x) - q_{m(X),\alpha}| \quad \text{for } \mathbf{P}_X\text{-almost all } x,$$

then the error of the Monte Carlo estimate $q^{(MC)}_{m_n(X),N_n,\alpha}$ of $q_{m(X),\alpha}$ is small, i.e.,

$$\left| q^{(MC)}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha} \right| = O_{\mathbf{P}}\left(\delta_n + \frac{1}{\sqrt{N_n}} \right),$$

where N_n is the sample size of the Monte Carlo estimate. We use this result to analyze the rate of convergence of two different estimates, whereas the error of m_n is globally small for the first estimate but only locally small for the second estimate. We show in particular that if m is (p, C)-smooth, i.e., if m is p-times continuously differentiable (see the exact definition below), then the first estimate achieves (up to some logarithmic factor) a rate of convergence of order $n^{-p/d}$ (as compared to the rate $n^{-1/2}$ of the order statistics estimate above), but the second one achieves (again up to some logarithmic factor) the rate of order $n^{-2p/d}$.

In order to construct the surrogate m_n any kind of nonparametric regression estimate can be used. Possible choices include kernel regression estimate (cf., e.g., Nadaraya [36, 37], Watson [50]), Devroye and Wagner [11], Stone [46, 47] or Devroye and Krzyżak [13]), partitioning regression estimate (cf., e.g., Györfi [20] or Beirlant and Györfi [2]), nearest neighbor regression estimate (cf., e.g., Devroye [12] or Devroye et al. [14]), orthogonal series regression estimate (cf., e.g., Rafajłowicz [42] or Greblicki and Pawlak [19]), least squares estimates (cf., e.g., Lugosi and Zeger [34] or Kohler [27]) or smoothing spline estimates (cf., e.g., Wahba [51] or Kohler and Krzyżak [28]). For survey of quantile regression we refer the reader to Yu et al. [52], Takeuchi et al. [48] and Koenker [26].

The idea of estimating the distribution of a random variable $m(X)$ by the distribution of $m_n(X)$, where m_n is a suitable surrogate (or estimate) of m, has been considered in a number of papers, see for example the discussion on applications of surrogate models in Sect. 2.

Various versions of importance sampling algorithms using surrogate models have been used in Dubourg et al. [15] and in Kohler et al. [31], whereas in the latter article theoretical results have also been provided.

4.1 A General Error Bound

Let X, X_1, X_2, ...be independent and identically distributed random variables. In this section we consider a general Monte Carlo surrogate quantile estimate defined as follows: First data

$$(x_1, m(x_1)), \ldots, (x_n, m(x_n))$$

is used to construct an estimate

$$m_n(\cdot) = m_n(\cdot, (x_1, m(x_1)), \dots, (x_n, m(x_n))) : \mathbb{R}^d \to \mathbb{R}$$

of m. Here $x_i = X_i$ is one possible choice for the values of $x_1, \dots, x_n \in \mathbb{R}^d$, but not the only one (see the next two sections below). Then $X_{n+1}, \dots, X_{n+N_n}$ are used to define a Monte Carlo estimate of the α-quantile of $m_n(X)$ by

$$\hat{q}^{(MC)}_{m_n(X),N_n,\alpha} = \inf \left\{ y \in \mathbb{R} \, : \, \hat{G}^{(MC)}_{m_n(X),N_n}(y) \geq \alpha \right\},$$

where

$$\hat{G}^{(MC)}_{m_n(X),N_n}(y) = \frac{1}{N_n} \sum_{i=1}^{N_n} I_{\{m_n(X_{n+i}) \leq y\}}.$$

Intuitively it is clear that the error of m_n will influence the error of the above quantile estimate. Our main result states that for the error of the above quantile estimate it is not important that the local error of m_n is small in areas where m is far away from the quantile to be estimated.

Theorem 3 *Let X be an $\mathbb{R}^d$-valued random variable, let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function and let $\alpha \in (0, 1)$. Define the Monte Carlo surrogate quantile estimate $\hat{q}^{(MC)}_{m_n(X),N_n,\alpha}$ of $q_{m(X),\alpha}$ as above and let $\hat{q}^{(MC)}_{m(X),N_n,\alpha}$ be the Monte Carlo quantile estimate of $q_{m(X),\alpha}$ based on $m(X_{n+1}), \dots, m(X_{n+N_n})$, i.e.,*

$$\hat{q}^{(MC)}_{m(X),N_n,\alpha} = \inf \left\{ y \in \mathbb{R} \, : \, \hat{G}^{(MC)}_{m(X),N_n}(y) \geq \alpha \right\},$$

where

$$\hat{G}^{(MC)}_{m(X),N_n}(y) = \frac{1}{N_n} \sum_{i=1}^{N_n} I_{\{m(X_{n+i}) \leq y\}}.$$

For $n \in \mathbb{N}$ let $\delta_n > 0$ be such that the estimate m_n satisfies

$$|m_n(X_{n+i}) - m(X_{n+i})| \leq \frac{\delta_n}{2} + \frac{1}{2} \cdot |q_{m(X),\alpha} - m(X_{n+i})| \quad \textit{for all } i \in \{1, \dots N_n\}. \tag{21}$$

Then we have

$$\left| \hat{q}^{(MC)}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha} \right| \leq \delta_n + 2 \cdot \left| \hat{q}^{(MC)}_{m(X),N_n,\alpha} - q_{m(X),\alpha} \right|.$$

For the proof of Theorem 3 we refer the reader to [17]. We immediately conclude from (4).

Corollary 2 *Let X be an $\mathbb{R}^d$-valued random variable, let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function and let $\alpha \in (0, 1)$. Assume that $m(X)$ has a density which* *is continuous and positive at $q_{m(X),\alpha}$. Define the Monte Carlo surrogate quantile estimate $\hat{q}^{(MC)}_{m_n(X),N_n,\alpha}$ of $q_{m(X),\alpha}$ as above. For $n \in \mathbb{N}$ let $\delta_n > 0$ be such that the estimate m_n satisfies (21) with probability one. Then*

$$\left|\hat{q}^{(MC)}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha}\right| = O_{\mathbf{P}}\left(\delta_n + \frac{1}{\sqrt{N_n}}\right)$$

Remark 2 If

$$|m_n(x) - m(x)| \leq \frac{\delta_n}{2} + \frac{1}{2} \cdot |m(x) - q_{m(X),\alpha}| \quad \text{for } \mathbf{P}_X\text{-almost all } x, \tag{22}$$

then (21) holds with probability one.

Remark 3 Condition (21) is in particular satisfied if we choose

$$\delta_n = 2 \cdot \|m_n - m\|_{\infty, supp(\mathbf{P}_X)},$$

so Corollary 2 implies

$$\left|\hat{q}^{(MC)}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha}\right| = O_{\mathbf{P}}\left(\|m_n - m\|_{\infty, supp(\mathbf{P}_X)} + \frac{1}{\sqrt{N_n}}\right). \tag{23}$$

Remark 4 If the support of X is unbounded, it might be difficult to construct estimates for which the error is uniformly small on the whole support as requested in Remark 2. But under suitable assumptions on the tails of X it suffices to approximate m on a compact set. Indeed, let $\beta_n > 0$ be such that

$$N_n \cdot \mathbf{P}\{X \notin [-\beta_n, \beta_n]^d\} \to 0 \quad (n \to \infty).$$

Then it follows from Theorem 3 that in this case

$$\left|\hat{q}^{(MC)}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha}\right| = O_{\mathbf{P}}\left(\|m_n - m\|_{\infty, supp(\mathbf{P}_X)\cap[-A_n,A_n]^d} + \frac{1}{\sqrt{N_n}}\right).$$

4.2 A Surrogate Quantile Estimate Based on a Non-adaptively Chosen Surrogate

In this section we choose m_n as a non-adaptively chosen spline approximand in the definition of our Monte Carlo surrogate quantile estimate.

To do this, we choose $\alpha > 0$ and set $\beta_n = \log(n)^\alpha$ and we define a spline approximand which approximates m on $[-\beta_n, \beta_n]^d$. To this end we use polynomial splines defined in Sect. 2 as follows.

Choose $K \in \mathbb{N}$ and $M \in \mathbb{N}_0$, and set $u_k = k \cdot \beta_n / K$ ($k \in \mathbb{Z}$). For $k \in \mathbb{Z}$ let $B_{k,M} : \mathbb{R} \to \mathbb{R}$ be the univariate B-spline of degree M with knot sequence $(u_l)_{l \in \mathbb{Z}}$ and support $supp(B_{k,M}) = [u_k, u_{k+M+1}]$. For more details on spline construction refer to Sect. 2.

Define $\mathscr{S}_{K,M}$ as the set of all linear combinations of all those tensor product B-splines, where the support has nonempty intersection with $[-\beta_n, \beta_n]^d$, i.e., we set

$$\mathscr{S}_{K,M} = \left\{ \sum_{\mathbf{k} \in \{-K-M, -K-M+1, \ldots, K-1\}^d} a_{\mathbf{k}} \cdot B_{\mathbf{k},M} \ : \ a_{\mathbf{k}} \in \mathbb{R} \right\}.$$

It can be shown by using standard arguments from spline theory, that the functions in $\mathscr{S}_{K,M}$ are in each component $(M-1)$-times continuously differentiable, that they are equal to a (multivariate) polynomial of degree less than or equal to M (in each component) on each rectangle

$$[u_{k_1}, u_{k_1+1}) \times \cdots \times [u_{k_d}, u_{k_d+1}) \quad (\mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{Z}^d), \tag{24}$$

and that they vanish outside of the set

$$\left[\beta_n - M \cdot \frac{\beta_n}{K}, \beta_n + M \cdot \frac{\beta_n}{K} \right]^d.$$

Next we define spline approximands using so-called quasi interpolands: For a function $m : [-\beta_n, \beta_n]^d \to \mathbb{R}$ we define an approximating spline by

$$(Qm)(x) = \sum_{\mathbf{k} \in \{-K-M, -K-M+1, \ldots, K-1\}^d} Q_{\mathbf{k}} m \cdot B_{\mathbf{k},M}$$

where

$$Q_{\mathbf{k}} m = \sum_{\mathbf{j} \in \{0,1,\ldots,M\}^d} a_{\mathbf{k},\mathbf{j}} \cdot m(t_{k_1,j_1}, \ldots, t_{k_d,j_d})$$

for some $a_{\mathbf{k},\mathbf{j}} \in \mathbb{R}$ defined below and for suitably chosen points $t_{k,j} \in supp(B_{k,M}) \cap [-\beta_n, \beta_n]$. It can be shown that if we set

$$t_{k,j} = \frac{k}{K} \cdot \beta_n + \frac{j}{K \cdot M} \cdot \beta_n = \frac{k \cdot M + j}{K \cdot M} \cdot \beta_n \quad (j \in \{0, \ldots, M\}, k \in \{-K, \ldots, K-1\})$$

and

$$t_{k,j} = -\beta_n + \frac{j}{K \cdot M} \quad (j \in \{0, \ldots, M\}, k \in \{-K-M, -K-M+1, \ldots, -K-1\}),$$

then there exist coefficients $a_{\mathbf{k},\mathbf{j}}$ (which can be computed by solving a linear equation system), such that

$$|Q_{\mathbf{k}} f| \leq c_1 \cdot \|f\|_{\infty,[u_{k_1},u_{k_1+M+1}]\times\cdots\times[u_{k_d},u_{k_d+M+1}]} \tag{25}$$

for any $\mathbf{k} \in \{-M, -M+1, \ldots, K-1\}^d$, any $f : [-\beta_n, \beta_n]^d \to \mathbb{R}$ and some universal constant c_1, and such that Q reproduces polynomials of degree M or less (in each component) on $[-\beta_n, \beta_n]^d$, i.e., for any multivariate polynomial $p : \mathbb{R}^d \to \mathbb{R}$ of degree M or less (in each component) we have

$$(Qp)(x) = p(x) \quad (x \in [-\beta_n, \beta_n]^d) \tag{26}$$

(cf., e.g., Theorems 14.4 and 15.2 in Györfi et al. [21]).

Next we define our estimate m_n as a quasi interpoland. We fix the degree $M \in \mathbb{N}$ and set

$$K = K_n = \left\lfloor \frac{\lfloor n^{1/d} \rfloor - 1}{2 \cdot M} \right\rfloor.$$

Furthermore we choose $x_1, \ldots, x_n$ such that all of the $(2 \cdot M \cdot K + 1)^d$ points of the form

$$\left(\frac{j_1}{M \cdot K} \cdot A_n, \ldots, \frac{j_d}{M \cdot K} \cdot A_n\right) \quad (j_1, \ldots, j_d \in \{-M \cdot K, -M \cdot K + 1, \ldots, M \cdot K\})$$

are contained in $\{x_1, \ldots, x_n\}$, which is possible since $(2 \cdot M \cdot K + 1)^d \leq n$. Then we define

$$m_n(x) = (Qm)(x),$$

where Qm is the above defined quasi interpoland satisfying (25) and (26). The computation of Qm requires only function values of m at the points x_1, …, x_n, i.e., the estimate depends on the data

$$(x_1, m(x_1)), \ldots, (x_n, m(x_n)),$$

and hence m_n is well defined.

Theorem 4 *Let X be an $\mathbb{R}^d$-valued random variable, let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function and let $\alpha \in (0, 1)$. Assume that $m(X)$ has a density which is continuous and positive at q_α and that m is (p, C)-smooth for some $p > 0$ and some $C > 0$. Define the Monte Carlo surrogate quantile estimate $\hat{q}^{(MC)}_{m_n(X),N_n,\alpha}$ of $q_{m(X),\alpha}$ as in Sect. 4.1, where m_n is the spline approximand defined above with parameter $M \geq p - 1$.*

Assume furthermore that

$$N_n \cdot \mathbf{P}\{X \notin [-\log(n)^\alpha, \log(n)^\alpha]^d\} \to 0 \quad (n \to \infty). \tag{27}$$

Then

$$\left|\hat{q}^{(MC)}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha}\right| = O_{\mathbf{P}}\left(\frac{\log(n)^{\alpha\cdot p}}{n^{p/d}} + \frac{1}{\sqrt{N_n}}\right).$$

In particular, if we set $N_n = \lceil n^{2p/d}/\log(n)^{2\cdot\alpha\cdot p}\rceil$ *then we get*

$$\left|\hat{q}^{(MC)}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha}\right| = O_{\mathbf{P}}\left(\frac{\log(n)^{\alpha\cdot p}}{n^{p/d}}\right).$$

It follows from Theorem 4 that in case that m be (p, C)-smooth for some $p > d/2$ or some $p > d$, respectively, and that

$$n^{2\cdot p/d}\cdot\mathbf{P}\{X \notin [-\log(n)^{\alpha}, \log(n)^{\alpha}]\} \to 0 \quad (n \to \infty),$$

then the above Monte Carlo surrogate quantile estimate achieves a rate of convergence better than $n^{-1/2}$ or n^{-1}, respectively. It follows from Markov inequality that (27) is for instance satisfied if

$$\mathbf{E}\{\exp(X)\} < \infty \quad \text{and} \quad \frac{N_n}{n^{\alpha}} \to 0 \quad (n \to \infty).$$

For in-depth discussion of the surrogate quantile estimate based on a non-adaptively and adaptively chosen surrogate the reader is referred to Enss et al. [17].

5 Conclusions

In the paper we discussed recent results on nonparametric quantile estimation using importance sampling, Robbins-Monro type importance sampling and surrogate models. We also dealt with nonparametric maximum quantile estimation. The problems discussed have both theoretical flavor and are important in applications. Further research on these problems is conducted.

Acknowledgments The author would like to thank his co-authors and acknowledge the support from the Natural Sciences and Engineering Research Council of Canada.

References

1. Arnold BC, Balakrishnan N, Nagaraja HN (1992) A first course in order statistics. Wiley, New York
2. Beirlant J, Györfi L (1998) On the asymptotic L_2-error in partitioning regression estimation. J Stat Plan Inference 71:93–107
3. Benveniste A, Métivier M, Priouret P (1990) Adaptive algorithms and stochastic approximation. Springer, New York

4. Bichon B, Eldred M, Swiler M, Mahadevan S, McFarland J (2008) Efficient global reliability analysis for nonlinear implicit performance functions. AIAA J 46:2459–2468
5. Bourinet JM, Deheeger F, Lemaire M (2011) Assessing small failure probabilities by combined subset simulation and support vector machines. Struct Saf 33:343–353
6. Cannamela C, Garnier J, Iooss B (2008) Controlled stratification for quantile estimation. Ann Appl Stat 2(4):1554–1580
7. Chen H-F (2002) Stochastic approximation and its applications. Kluwer Academic Publishers, Boston
8. Das PK, Zheng Y (2000) Cumulative formation of response surface and its use in reliability analysis. Probab Eng Mech 15:309–315
9. de Boor C (1978) A practical guide to splines. Springer, New York
10. Deheeger F, Lemaire M (2010) Support vector machines for efficient subset simulations: ^{2}SMART method. In: Proceedings of the 10th international conference on applications of statistics and probability in civil engineering (ICASP10), Tokyo, Japan
11. Devroye L, Wagner TJ (1980) Distribution-free consistency results in nonparametric discrimination and regression function estimation. Ann Stat 8:231–239
12. Devroye L (1982) Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 61:467–481
13. Devroye L, Krzyżak A (1989) An equivalence theorem for L_1 convergence of the kernel regression estimate. J Stat Plan Inference 23:71–82
14. Devroye L, Györfi L, Krzyżak A, Lugosi G (1994) On the strong universal consistency of nearest neighbor regression function estimates. Ann Stat 22:1371–1385
15. Dubourg V, Sudret B, Deheeger F (2013) Metamodel-based importance sampling for structural reliability analysis. Probab Eng Mech 33:47–57
16. Egloff D, Leippold M (2010) Quantile estimation with adaptive importance sampling. Ann Stat 38(2):1244–1278
17. Enss GC, Kohler M, Krzyżak A, Platz R (2014) Nonparametric quantile estimation based on surrogate models. Submitted for publication
18. Glasserman P (2004) Monte Carlo methods in financial engineering. Springer, New York
19. Greblicki W, Pawlak M (1985) Fourier and Hermite series estimates of regression functions. Ann Inst Stat Math 37:443–454
20. Györfi L (1981) Recent results on nonparametric regression estimate and multiple classification. Probl Control Inf Theory 10:43–52
21. Györfi L, Kohler M, Krzyżak A, Walk H (2002) A distribution-free theory of nonparametric regression. Springer series in statistics. Springer, New York
22. Holst U (1987) Recursive estimation of quantiles using recursive kernel density estimators. Seq Anal: Des Methods Appl 6(3):219–237
23. Hurtado J (2004) Structural reliability—statistical learning perspectives, Lecture notes in applied and computational mechanics, vol 17. Springer, New York
24. Kaymaz I (2005) Application of Kriging method to structural reliability problems. Struct Saf 27:133–151
25. Kim SH, Na SW (1997) Response surface method using vector projected sampling points. Struct Saf 19:3–19
26. Koenker R (2005) Quantile regression. Cambridge University Press, New York
27. Kohler M (2000) Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. J Stat Plan Inference 89:1–23
28. Kohler M, Krzyżak A (2001) Nonparametric regression estimation using penalized least squares. IEEE Trans Inf Theory 47:3054–3058
29. Kohler M (2014) Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design. J Multivar Anal 132:197–208
30. Kohler M, Krzyżak A, Walk H (2014) Nonparametric recursive quantile estimation. Stat Probab Lett 93:102–107

31. Kohler M, Krzyżak A, Tent R, Walk H (2014) Nonparametric quantile estimation using importance sampling. Submitted for publication
32. Kushner HJ, Yin G (2003) Stochastic approximation and recursive algorithms and applications, 2nd edn. Springer, New York
33. Ljung L, Pflug G, Walk H (1992) Stochastic approximation and optimization of random systems. Birkhäuser Verlag, Basel
34. Lugosi G, Zeger K (1995) Nonparametric estimation via empirical risk minimization. IEEE Trans Inf Theory 41:677–687
35. Morio J (2012) Extreme quantile estimation with nonparametric adaptive importance sampling. Simul Model Pract Theory 27:76–89
36. Nadaraya EA (1964) On estimating regression. Theory Probab Appl 9:141–142
37. Nadaraya EA (1970) Remarks on nonparametric estimates for density functions and regression curves. Theory Probab Appl 15:134–137
38. Neddermeyer JC (2009) Computationally efficient nonparametric importance sampling. J Am Stat Assoc 104(486):788–802
39. Oakley J (2004) Estimating percentiles of uncertain computer code outputs. J R Stat Soc: Ser C (Appl Stat) 53(1):83–93
40. Papadrakakis M, Lagaros N (2002) Reliability-based structural optimization using neural networks and Monte Carlo simulation. Comput Methods Appl Mech Eng 191:3491–3507
41. Polyak BT, Juditsky AB (2002) Acceleration of stochastic approximation by averaging. SIAM J Control Optim 30(4):838–855
42. Rafajłowicz E (1987) Nonparametric orthogonal series estimators of regression: a class attaining the optimal convergence rate in L2. Stat Probab Lett 5:219–224
43. Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22(3):400–407
44. Ruppert D (1991) Stochastic approximation. In: Gosh BK, Sen PK (eds) Handbook of sequential analysis, Ch. 22. Marcel Dekker, New York, pp 503–529
45. Santner TJ, Williams BJ, Notz WI (2003) The design and analysis of computer experiments. Springer, New York
46. Stone CJ (1977) Consistent nonparametric regression. Ann Stat 5:595–645
47. Stone CJ (1982) Optimal global rates of convergence for nonparametric regression. Ann Stat 10:1040–1053
48. Takeuchi I, Le QV, Sears TD, Smola AJ (2006) Nonparametric quantile estimation. J Mach Learn Res 7:1231–1264
49. Tierney L (1983) A space-efficient recursive procedure for estimating a quantile of an unknown distribution. SIAM J Sci Stat Comput 4(4):706–711
50. Watson GS (1964) Smooth regression analysis. Sankhya Ser A 26:359–372
51. Wahba G (1990) Spline models for observational data. SIAM, Philadelphia
52. Yu K, Lu Z, Stander J (2003) Quantile regression: application and current research areas. J R Stat Soc, Ser D 52:331–350

Adaptive Monte Carlo Maximum Likelihood

Błażej Miasojedow, Wojciech Niemiro, Jan Palczewski and Wojciech Rejchel

Abstract We consider Monte Carlo approximations to the maximum likelihood estimator in models with intractable norming constants. This paper deals with adaptive Monte Carlo algorithms, which adjust control parameters in the course of simulation. We examine asymptotics of adaptive importance sampling and a new algorithm, which uses resampling and MCMC. This algorithm is designed to reduce problems with degeneracy of importance weights. Our analysis is based on martingale limit theorems. We also describe how adaptive maximization algorithms of Newton-Raphson type can be combined with the resampling techniques. The paper includes results of a small scale simulation study in which we compare the performance of adaptive and non-adaptive Monte Carlo maximum likelihood algorithms.

Keywords Maximum likelihood · Importance sampling · Adaptation · MCMC · Resampling

B. Miasojedow · W. Niemiro (✉)
Faculty of Mathematics, Informatics and Mechanics,, University of Warsaw,
Warsaw, Poland
e-mail: W.Niemiro@mimuw.edu.pl

B. Miasojedow
e-mail: W.Miasojedow@mimuw.edu.pl

J. Palczewski
School of Mathematics, University of Leeds, Leeds, UK
e-mail: J.Palczewski@leeds.ac.uk

W. Niemiro · W. Rejchel
Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland
e-mail: wrejchel@gmail.com

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_14

1 Introduction

Maximum likelihood (ML) is a well-known and often used method in estimation of parameters in statistical models. However, for many complex models exact calculation of such estimators is very difficult or impossible. Such problems arise if considered densities are known only up to intractable norming constants, for instance in Markov random fields or spatial statistics. The wide range of applications of models with unknown norming constants is discussed e.g. in [10]. Methods proposed to overcome the problems with computing ML estimates in such models include, among others, maximum pseudolikelihood [2], "coding method" [9] and Monte Carlo maximum likelihood (MCML) [5, 9, 15, 17]. In our paper we focus on MCML.

In influential papers [4, 5] the authors prove consistency and asymptotic normality of MCML estimators. To improve the performance of MCML, one can adjust control parameters in the course of simulation. This leads to adaptive MCML algorithms. We generalize the results of the last mentioned papers first to an adaptive version of importance sampling and then to a more complicated adaptive algorithm which uses resampling and Markov chain Monte Carlo (MCMC) [7]. Our analysis is asymptotic and it is based on the martingale structure of the estimates. The main motivating examples are the autologistic model (with or without covariates) and its applications to spatial statistics as described e.g. in [9] and the autonormal model [11].

2 Adaptive Importance Sampling

Denote by $f_\theta, \theta \in \Theta$, a family of unnormalized densities on space $\mathcal{Y}$. A dominating measure with respect to which these densities are defined is denoted for simplicity by $\mathrm{d}y$. Let y_{obs} be an observation. We intend to find the maximizer $\theta_\star$ of the log-likelihood

$$\ell(\theta) = \log f_\theta(y_{\mathrm{obs}}) - \log c(\theta),$$

where $c(\theta)$ is the normalizing constant. We consider the situation where this constant,

$$c(\theta) = \int_{\mathcal{Y}} f_\theta(y)\mathrm{d}y,$$

is *unknown and numerically intractable*. It is approximated with Monte Carlo simulation, resulting in

$$\ell_m(\theta) = \log f_\theta(y_{\mathrm{obs}}) - \log \hat{c}_m(\theta), \tag{2.1}$$

where $\hat{c}_m(\theta)$ is a Monte Carlo (MC) estimate of $c(\theta)$. The classical importance sampling (IS) estimate is of the form

$$\hat{c}_m(\theta) = \frac{1}{m}\sum_{j=1}^{m}\frac{f_\theta(Y_j)}{h(Y_j)}, \tag{2.2}$$

where $Y_1, \ldots, Y_m$ are i.i.d. samples from an instrumental density h. Clearly, an optimal choice of h depends on the maximizer $\theta_\star$ of ℓ, so we should be able to improve our initial guess about h while the simulation progresses. This is the idea behind *adaptive importance sampling* (AIS). A discussion on the choice of instrumental density is deferred to subsequent subsections. Let us describe an adaptive algorithm in the following form, suitable for further generalizations. Consider a parametric family h_ψ, $\psi \in \Psi$ of instrumental densities.

```
Algorithm AdapIS

1. Set an initial value of $\psi_1, m = 1, \hat{c}_0(\theta) \equiv 0$.
2. Draw $Y_m \sim h_{\psi_m}$.
3. Update the approximation of $c(\theta)$:

   $$\hat{c}_m(\theta) = \frac{m-1}{m}\hat{c}_{m-1}(\theta) + \frac{1}{m}\frac{f_\theta(Y_m)}{h_{\psi_m}(Y_m)}.$$

4. Update $\psi$: choose $\psi_{m+1}$ based on the history of the
   simulation.
5. $m = m + 1$; go to 2.
```

At the output of this algorithm we obtain an AIS estimate

$$\hat{c}_m(\theta) = \frac{1}{m}\sum_{j=1}^{m}\frac{f_\theta(Y_j)}{h_{\psi_j}(Y_j)}. \tag{2.3}$$

The samples Y_j are neither independent nor have the same distribution. However (2.3) has a nice *martingale* structure. If we put

$$\mathcal{F}_m = \sigma\left\{Y_j, \psi_j : j \leq m\right\}$$

then ψ_{m+1} is $\mathcal{F}_m$-measurable. The well-known property of unbiasedness of IS implies that

$$\mathbb{E}\left(\frac{f_\theta(Y_{m+1})}{h_{\psi_{m+1}}(Y_{m+1})}\middle|\mathcal{F}_m\right) = c(\theta). \tag{2.4}$$

In other words, $f_\theta(Y_m)/h_{\psi_m}(Y_m) - c(\theta)$ are martingale differences (MGD), for every fixed θ.

2.1 Hypo-convergence of ℓ_m and Consistency of $\hat{\theta}_m$

In this subsection we make the following assumptions.

Assumption 1 For any $\theta \in \Theta$

$$\sup_{\psi} \int \frac{f_\theta(y)^2}{h_\psi(y)} \mathrm{d}y < \infty.$$

Assumption 2 The mapping $\theta \mapsto f_\theta(y)$ is continuous for each fixed y.

Assumption 1 implies that for any θ, there is a constant $M_\theta < \infty$ such that for all j,

$$\mathbb{E}\left(\left(\frac{f_\theta(Y_j)}{h_{\psi_j}(Y_j)}\right)^2 \middle| \mathcal{F}_{j-1}\right) \leq M_\theta, \quad \text{a.s.},$$

because $Y_j \sim h_{\psi_j}$. Note that Assumption 1 is trivially true if the mapping $y \mapsto f_\theta(y)/h_\psi(y)$ is uniformly bounded for $\theta \in \Theta$, $\psi \in \Psi$. Recall also that

$$m(\hat{c}_m(\theta) - c(\theta)) = \sum_{j=1}^{m} \left(\frac{f_\theta(Y_j)}{h_{\psi_j}(Y_j)} - c(\theta)\right)$$

is a zero-mean martingale. Under Assumption 1, for a fixed $\theta \in \Theta$, we have $\hat{c}_m(\theta) \to c(\theta)$ a.s. by the SLLN for martingales (see Theorem A.2, Appendix A), so $\ell_m(\theta) \to \ell(\theta)$ a.s. This is, however, insufficient to guarantee the convergence of maximum likelihood estimates $\hat{\theta}_m$ (maximizers of ℓ_m) to $\theta_\star$. Under our assumptions we can prove hypo-convergence of the log-likelihood approximations.

Definition 1 A sequence of functions g_m epi-converges to g if for any x we have

$$\sup_{B \in N(x)} \limsup_{m\to\infty} \inf_{y\in B} g_m(y) \leq g(x),$$
$$\sup_{B \in N(x)} \liminf_{m\to\infty} \inf_{y\in B} g_m(y) \geq g(x),$$

where $N(x)$ is a family of all (open) neighbourhoods of x.

A sequence of functions g_m hypo-converges to g if $(-g_m)$ epi-converges to $(-g)$.

An equivalent definition of epi-convergence is in the following theorem:

Theorem 1 ([14, Proposition 7.2]) *g_m epi-converges to g iff at every point x*

$$\limsup_{m\to\infty} g_m(x_m) \leq g(x), \qquad \textit{for some sequence } x_m \to x,$$
$$\liminf_{m\to\infty} g_m(x_m) \geq g(x), \qquad \textit{for every sequence } x_m \to x.$$

As a corollary to this theorem comes the proposition that will be used to prove convergence of $\hat{\theta}_m$, the maximizer of ℓ_m, to $\theta_\star$ (see, also, [1, Theorem 1.10]).

Proposition 1 *Assume that g_m epi-converges to g, $x_m \to x$ and $g_m(x_m) - \inf g_m \to 0$. Then $g(x) = \inf_y g(y) = \lim_{m\to\infty} g_m(x_m)$.*

Proof (We will use Theorem 1 many times.) Let y_m be a sequence converging to x and such that $\limsup_{m\to\infty} g_m(y_m) \le g(x)$ (such sequence y_m exists). This implies that $\limsup_{m\to\infty} \inf g_m \le g(x)$. On the other hand, $g(x) \le \liminf_{m\to\infty} g_m(x_m) = \liminf_{m\to\infty} \inf g_m$, where the equality follows from the second assumption on x_m. Summarizing, $g(x) = \lim_{m\to\infty} \inf g_m = \lim_{m\to\infty} g_m(x_m)$. In particular, $\inf g \le \lim_{m\to\infty} \inf g_m$.

Take any $\varepsilon > 0$ and let x_ε be such that $g(x_\varepsilon) \le \inf g + \varepsilon$. There exists a sequence y_m converging to x_ε such that $g(x_\varepsilon) \ge \limsup_{m\to\infty} g_m(y_m) \ge \limsup_{m\to\infty} \inf g_m$, hence $\lim_{m\to\infty} \inf g_m \le \inf g + \varepsilon$. By arbitrariness of $\varepsilon > 0$ we obtain $\lim_{m\to\infty} \inf g_m \le \inf g$. This completes the proof.

Theorem 2 *If Assumptions 1 and 2 are satisfied, then ℓ_m hypo-converges to ℓ almost surely.*

Proof The proof is similar to the proof of Theorem 1 in [4]. We have to prove that $\hat{c}_m$ epi-converges to c. Fix $\theta \in \Theta$.

Step 1: For any $B \in N(\theta)$, we have

$$\liminf_{m\to\infty} \inf_{\varphi\in B} \hat{c}_m(\varphi) \ge \int \inf_{\varphi\in B} f_\varphi(y)\mathrm{d}y =: \underline{c}(B). \tag{2.5}$$

Indeed,

$$\begin{aligned}\inf_{\varphi\in B} \hat{c}_m(\phi) &= \inf_{\varphi\in B} \frac{1}{m}\sum_{j=1}^m \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)} \ge \frac{1}{m}\sum_{j=1}^m \inf_{\varphi\in B} \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)} \\ &= \frac{1}{m}\sum_{j=1}^m \Big(\inf_{\varphi\in B} \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)} - \underline{c}(B)\Big) + \underline{c}(B).\end{aligned}$$

The sum is that of martingale differences, so assuming that there is $M < \infty$ such that

$$\sup_j \mathbb{E}\bigg(\Big(\inf_{\varphi\in B} \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)} - \underline{c}(B)\Big)^2 \bigg| \mathcal{F}_{j-1}\bigg) \le M$$

the SLLN implies (2.5). We have the following estimates:

$$\mathbb{E}\left(\left(\inf_{\varphi\in B}\frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)}-\underline{c}(B)\right)^2\middle|\mathcal{F}_{j-1}\right)=\mathrm{Var}\left(\inf_{\varphi\in B}\frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)}\middle|\mathcal{F}_{j-1}\right)$$
$$\leq\mathbb{E}\left(\left(\inf_{\varphi\in B}\frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)}\right)^2\middle|\mathcal{F}_{j-1}\right)\leq\mathbb{E}\left(\left(\frac{f_\theta(Y_j)}{h_{\psi_j}(Y_j)}\right)^2\middle|\mathcal{F}_{j-1}\right)\leq M_\theta,$$

where the last inequality is by Assumption 1.

Step 2: We shall prove that $\sup_{B\in N(\theta)}\liminf_{m\to\infty}\inf_{\varphi\in B}\hat{c}_m(\phi)\geq c(\theta)$.

The left-hand side is bounded from below by $\sup_{B\in N(\theta)}\underline{c}(B)$. Further, we have

$$\sup_{B\in N(\theta)}\underline{c}(B)\geq\lim_{\delta\downarrow 0}\underline{c}(B(\theta,\delta))=\int\lim_{\delta\downarrow 0}\inf_{\varphi\in B(\theta,\delta)}f_\varphi(y)\mathrm{d}y=\int f_\theta(y)\mathrm{d}y=c(\theta),$$

where the first equality follows from the dominated convergence theorem (the dominator is f_θ) and the last—from the Assumption 2.

Step 3: We have

$$\sup_{B\in N(\theta)}\limsup_{m\to\infty}\inf_{\varphi\in B}\hat{c}_m(\varphi)\leq\sup_{B\in N(\theta)}\inf_{\varphi\in B}\limsup_{m\to\infty}\hat{c}_m(\varphi)=\sup_{B\in N(\theta)}\inf_{\varphi\in B}c(\varphi)\leq c(\theta).$$

Hence, $\sup_{B\in N(\theta)}\limsup_{m\to\infty}\inf_{\varphi\in B}\hat{c}_m(\varphi)\leq c(\theta)$.

Note that almost sure convergence in the next Proposition corresponds to the randomness introduced by `AdapIS` and y_{obs} is fixed throughout this paper.

Proposition 2 *If Assumptions 1 and 2 hold, $\theta_\star$ is the unique maximizer of ℓ and sequence $(\hat{\theta}_m)$ (where $\hat{\theta}_m$ maximizes ℓ_m) is almost surely bounded then $\hat{\theta}_m\to\theta_\star$ almost surely.*

Proof As we have already mentioned, by SLLN for martingales, $\ell_m(\theta)\to\ell(\theta)$, pointwise. Hypo-convergence of ℓ_m to ℓ implies, by Proposition 1, that the maximizers of ℓ_m have accumulation points that are the maximizers of ℓ. If ℓ has a unique maximizer $\theta_\star$ then any convergent subsequence of $\hat{\theta}_m$, maximizers of ℓ_m, converges to $\theta_\star$. The conclusion follows immediately.

Of course, it is not easy to show boundedness of $\hat{\theta}_m$ in concrete examples. In the next section we will prove consistency of $\hat{\theta}_m$ in models where log-likelihoods and their estimates are concave.

2.2 Central Limit Theorem for Adaptive Importance Sampling

Let $\hat{\theta}_m$ be a maximizer of ℓ_m, i.e. the AIS estimate of the likelihood given by (2.1) with (2.3). We assume that $\theta_\star$ is a unique maximizer of ℓ. For asymptotic normality of $\hat{\theta}_m$, we will need the following assumptions.

Assumption 3 First and second order derivatives of f_θ with respect to θ (denoted by ∇f_θ and $\nabla^2 f_\theta$) exist in a neighbourhood of $\theta_\star$ and we have

$$\nabla c(\theta) = \int \nabla f_\theta(y)\mathrm{d}y, \qquad \nabla^2 c(\theta) = \int \nabla^2 f_\theta(y)\mathrm{d}y.$$

Assumption 4 $\hat{\theta}_m = \theta_\star + O_\mathrm{p}(1/\sqrt{m})$.

Assumption 5 Matrix $D = \nabla^2 \ell(\theta_\star)$ is negative definite.

Assumption 6 For every y, function $\psi \mapsto h_\psi(y)$ is continuous and $h_\psi(y) > 0$.

Assumption 7 For some $\psi_\star$ we have $\psi_m \to \psi_\star$ almost surely.

Assumption 8 There exists a nonnegative function g such that $\int g(y)\mathrm{d}y < \infty$ and the inequalities

$$\sup_\psi \frac{f_{\theta_\star}(y)^{2+\alpha}}{h_\psi(y)^{1+\alpha}} \le g(y), \quad \sup_\psi \frac{|\nabla f_{\theta_\star}(y)|^{2+\alpha}}{h_\psi(y)^{1+\alpha}} \le g(y),$$
$$\sup_\psi \frac{\|\nabla^2 f_{\theta_\star}(y)\|^{1+\alpha}}{h_\psi(y)^{\alpha}} \le g(y)$$

are fulfilled for some $\alpha > 0$ and also for $\alpha = 0$.

Assumption 9 Functions $\nabla^2 \ell_m(\theta)$ are asymptotically stochastically equicontionuous at $\theta_\star$, i.e. for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\limsup_{m\to\infty} \mathbb{P}\left(\sup_{|\theta-\theta_\star|\le\delta} \|\nabla^2 \ell_m(\theta) - \nabla^2 \ell_m(\theta_\star)\| > \varepsilon \right) = 0.$$

Let us begin with some comments on these assumptions and note simple facts which follow from them. Assumption 3 is a standard regularity condition. It implies that a martingale property similar to (2.4) holds also for the gradients and hessians:

$$\mathbb{E}\left(\frac{\nabla f_\theta(Y_{m+1})}{h_{\psi_{m+1}}(Y_{m+1})} \middle| \mathcal{F}_m \right) = \nabla c(\theta), \quad \mathbb{E}\left(\frac{\nabla^2 f_\theta(Y_{m+1})}{h_{\psi_{m+1}}(Y_{m+1})} \middle| \mathcal{F}_m \right) = \nabla^2 c(\theta). \qquad (2.6)$$

Assumption 4 stipulates square root consistency of $\hat{\theta}_m$. It is automatically fulfilled if ℓ_m is concave, in particular for exponential families. Assumption 7 combined with 6 is a "diminishing adaptation" condition. It may be ensured by an appropriately specifying step 4 of `AdapIS`. The next assumptions are not easy to verify in general, but they are satisfied for exponential families on finite spaces, in particular for our "motivating example": autologistic model. Let us also note that our Assumption 9 plays a similar role to Assumption (f) in [4, Theorem 7].

Assumption 8 together with (2.4) and (2.6) allows us to apply SLLN for martingales in a form given in Theorem A.2, Appendix A. Indeed, $f_{\theta_\star}(Y_m)/h_{\psi_m}(Y_m) - c(\theta_\star)$, $\nabla f_{\theta_\star}(Y_m)/h_{\psi_m}(Y_m) - \nabla c(\theta_\star)$ and $\nabla^2 f_{\theta_\star}(Y_m)/h_{\psi_m}(Y_m) - \nabla^2 c(\theta_\star)$ are MGDs with bounded moments of order $1+\alpha > 1$. It follows that, almost surely,

$$\hat{c}_m(\theta_\star) \to c(\theta_\star), \qquad \nabla\hat{c}_m(\theta_\star) \to \nabla c(\theta_\star), \qquad \nabla^2\hat{c}_m(\theta_\star) \to \nabla^2 c(\theta_\star). \tag{2.7}$$

Now we are in a position to state the main result of this section.

Theorem 3 *If Assumptions 3–9 hold then*

$$\sqrt{m}\left(\hat{\theta}_m - \theta_\star\right) \to \mathcal{N}(0, D^{-1} V D^{-1}) \quad \textit{in distribution},$$

where $D = \nabla^2 \ell(\theta_\star)$ *and*

$$V = \frac{1}{c(\theta_\star)^2} \mathrm{VAR}_{Y \sim h_{\psi_\star}} \left[\frac{\nabla f_{\theta_\star}(Y)}{h_{\psi_\star}(Y)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{f_{\theta_\star}(Y)}{h_{\psi_\star}(Y)} \right],$$

where $\psi_\star$ *is defined in Assumption 7.*

Proof It is well-known (see [12, Theorem VII.5]) that we need to prove

$$\sqrt{m}\nabla\ell_m(\theta_\star) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V) \tag{2.8}$$

and that for every $M > 0$, the following holds:

$$\begin{aligned} \sup_{|\theta-\theta_\star| \le M/\sqrt{m}} m \Big| & \ell_m(\theta) - \ell_m(\theta_\star) \\ & - (\theta - \theta_\star)^\top \nabla\ell_m(\theta_\star) - \frac{1}{2}(\theta-\theta_\star)^\top D(\theta-\theta_\star) \Big| \xrightarrow{\mathrm{p}} 0. \end{aligned} \tag{2.9}$$

First we show (2.8). Since $\nabla\ell_m(\theta) = \nabla f_\theta(y_{\mathrm{obs}})/f_\theta(y_{\mathrm{obs}}) - \nabla\hat{c}_m(\theta)/\hat{c}_m(\theta)$ and $\nabla\ell(\theta_\star) = \nabla f_{\theta_\star}(y_{\mathrm{obs}})/f_{\theta_\star}(y_{\mathrm{obs}}) - \nabla c(\theta_\star)/c(\theta_\star) = 0$, we obtain that

$$\nabla\ell_m(\theta_\star) = \frac{\nabla c(\theta_\star)}{c(\theta_\star)} - \frac{\nabla\hat{c}_m(\theta_\star)}{\hat{c}_m(\theta_\star)} = \frac{\frac{\nabla c(\theta_\star)}{c(\theta_\star)}\hat{c}_m(\theta_\star) - \nabla\hat{c}_m(\theta_\star)}{\hat{c}_m(\theta_\star)}. \tag{2.10}$$

The denominator in the above expression converges to $c(\theta_\star)$ in probability, by (2.7). In view of Slutski's theorem, to prove (2.8) it is enough to show asymptotic normality of the numerator. We can write

$$\frac{\nabla c(\theta_\star)}{c(\theta_\star)}\hat{c}_m(\theta_\star) - \nabla\hat{c}_m(\theta_\star) = -\frac{1}{m}\sum_{j=1}^{m} \xi_j,$$

where we use the notation

$$\xi_j = \frac{\nabla f_{\theta_\star}(Y_j)}{h_{\psi_j}(Y_j)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{f_{\theta_\star}(Y_j)}{h_{\psi_j}(Y_j)}.$$

Now note that ξ_j are martingale differences by (2.4) and (2.6). Moreover,

$$\mathbb{E}\Big(\xi_j \xi_j^T | \mathcal{F}_{j-1}\Big) = \int \left(\frac{\nabla f_{\theta_\star}(y)}{h_{\psi_j}(y)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{f_{\theta_\star}(y)}{h_{\psi_j}(y)} \right) \left(\frac{\nabla f_{\theta_\star}(y)}{h_{\psi_j}(y)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{f_{\theta_\star}(y)}{h_{\psi_j}(y)} \right)^\top h_{\psi_j}(y) \mathrm{d}y,$$

so Assumptions 6 and 7 via dominated convergence and Assumption 8 (with $\alpha = 0$ in the exponent) entail

$$\mathbb{E}\Big(\xi_j \xi_j^T | \mathcal{F}_{j-1}\Big) \xrightarrow{\text{a.s.}} c(\theta_\star)^2 V.$$

Now we use Assumption 8 (with $\alpha > 0$ in the exponent) to infer the Lyapunov-type condition

$$\mathbb{E}\Big(|\xi_j|^{2+\alpha} | \mathcal{F}_{j-1}\Big) \le \text{const} \cdot \int g(y)\mathrm{d}y < \infty.$$

The last two displayed formulas are sufficient for a martingale CLT (Theorem A.1, Appendix A). We conclude that

$$\frac{1}{\sqrt{m}} \sum_{j=1}^{m} \xi_j \xi_j{}^\top \xrightarrow{\mathrm{d}} \mathcal{N}(0, c(\theta_\star)^2 V),$$

hence the proof of (2.8) is complete.

Now we proceed to a proof of (2.9). By Taylor expansion,

$$\ell_m(\theta) - \ell_m(\theta_\star) - (\theta - \theta_\star)^\top \nabla \ell_m(\theta_\star) = \frac{1}{2} (\theta - \theta_\star)^\top \nabla^2 \ell_m(\tilde{\theta})(\theta - \theta_\star)$$

for some $\tilde{\theta} \in [\theta, \theta_\star]$. Consequently, the LHS of (2.9) is

$$\begin{aligned}
&\le \sup_{\substack{|\theta - \theta_\star| \le M/\sqrt{m} \\ \tilde{\theta} \in [\theta, \theta_\star]}} m \left| \frac{1}{2} (\theta - \theta_\star)^\top \Big(\nabla^2 \ell_m(\tilde{\theta}) - \nabla^2 \ell(\theta_\star) \Big) (\theta - \theta_\star) \right| \\
&\le \frac{M^2}{2} \sup_{|\theta - \theta_\star| \le M/\sqrt{m}} \left\| \nabla^2 \ell_m(\theta) - \nabla^2 \ell(\theta_\star) \right\| \\
&\le \frac{M^2}{2} \sup_{|\theta - \theta_\star| \le M/\sqrt{m}} \left\| \nabla^2 \ell_m(\theta) - \nabla^2 \ell_m(\theta_\star) \right\| + \frac{M^2}{2} \left\| \nabla^2 \ell_m(\theta_\star) - \nabla^2 \ell(\theta_\star) \right\|.
\end{aligned}$$

The first term above goes to zero in probability by Assumption 9. The second term also goes to zero because

$$
\begin{aligned}
\nabla^2 \ell_m(\theta_\star) - \nabla^2 \ell(\theta_\star) &= \nabla^2 \log c(\theta_\star) - \nabla^2 \log \hat{c}_m(\theta_\star) \\
&= \frac{\nabla^2 c(\theta_\star)}{c(\theta_\star)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{\nabla c(\theta_\star)^\top}{c(\theta_\star)} - \frac{\nabla^2 \hat{c}_m(\theta_\star)}{\hat{c}_m(\theta_\star)} + \frac{\nabla \hat{c}_m(\theta_\star)}{\hat{c}_m(\theta_\star)} \frac{\nabla \hat{c}_m(\theta_\star)^\top}{\hat{c}_m(\theta_\star)} \xrightarrow{\mathrm{p}} 0,
\end{aligned}
$$

in view of (2.7). Therefore (2.9) holds and the proof is complete.

2.3 Optimal Importance Distribution

We advocate adaptation to improve the choice of instrumental distribution h. But which h is the best? If we use (non-adaptive) importance sampling with instrumental distribution h then the maximizer $\hat{\theta}_m$ of the MC likelihood approximation has asymptotic normal distribution, namely $\sqrt{m}\left(\hat{\theta}_m - \theta_\star\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, D^{-1} V D^{-1})$, $(m \to \infty)$ with

$$
V = \frac{1}{c(\theta_\star)^2} \mathrm{VAR}_{Y \sim h} \left[\frac{\nabla f_{\theta_\star}(Y)}{h(Y)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{f_{\theta_\star}(Y)}{h(Y)} \right].
$$

This fact is well-known [4] and is a special case of Theorem 3. Since the asymptotic distribution is multidimensional its dispersion can be measured in various ways, e.g., though the determinant, the maximum eigenvalue or the trace of the covariance matrix. We examine the trace which equals to the asymptotic mean square error of the MCML approximation (the asymptotic bias is nil). Notice that

$$
c(\theta_\star)^2 V = \mathrm{VAR}_{Y \sim h} \frac{\eta(Y)}{h(Y)} = \mathbb{E}_{Y \sim h} \frac{\eta(Y)\eta(Y)^\top}{h(Y)^2} = \int \frac{\eta(y)\eta(y)^\top}{h(y)} \mathrm{d}y,
$$

where

$$
\eta(y) = \nabla f_{\theta_\star}(y) - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} f_{\theta_\star}(y).
$$

Since $\mathrm{tr}\left[D^{-1}\eta(y)\eta(y)^\top D^{-1}\right] = (D^{-1}\eta(y))^\top D^{-1}\eta(y) = |D^{-1}\eta(y)|^2$, the minimization of $\mathrm{tr}(D^{-1} V D^{-1})$ is equivalent to

$$
\int \frac{|D^{-1}\eta(y)|^2}{h(y)} \mathrm{d}y \to \min,
$$

subject to $h \geq 0$ and $\int h = 1$. By Schwarz inequality we have

$$\int \frac{|D^{-1}\eta(y)|^2}{h(y)}\mathrm{d}y = \int \left(\frac{|D^{-1}\eta(y)|}{\sqrt{h(y)}}\right)^2 \mathrm{d}y \int \left(\sqrt{h(y)}\right)^2 \mathrm{d}y$$
$$\geq \left(\int \frac{|D^{-1}\eta(y)|}{\sqrt{h(y)}}\sqrt{h(y)}\mathrm{d}y\right)^2 = \left(\int |D^{-1}\eta(y)|\mathrm{d}y\right)^2,$$

with equality only for $h(y) \propto |D^{-1}\eta(y)|$. The optimum choice of h is therefore

$$h_\star(y) \propto \left|D^{-1}\left(\nabla f_{\theta_\star}(y) - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} f_{\theta_\star}(y)\right)\right|. \tag{2.11}$$

Unfortunately, this optimality result is chiefly of theoretical importance, because it is not clear how to sample from $h_\star$ and how to compute the norming constant for this distribution. This might well be even more difficult than computing $c(\theta)$.

The following example shows some intuitive meaning of (2.11). It is a purely "toy example" because the simple analitical formulas exist for $c(\theta)$ and $\theta_\star$ while MC is considered only for illustration.

Example 1 Consider a binomial distribution on $\mathcal{Y} = \{0, 1, \ldots, n\}$ given by $\pi_\theta(y) = \binom{n}{y} p^y (1-p)^{n-y}$. Parametrize the model with the log-odds-ratio $\theta = \log p/(1-p)$, absorb the $\binom{n}{y}$ factor into the measure $\mathrm{d}y$ to get the standard exponential family form with

$$f_\theta(y) = \mathrm{e}^{\theta y} \quad \text{and} \quad c(\theta) = \sum_{y=0}^{n} \binom{n}{y} \mathrm{e}^{\theta y} = (1+\mathrm{e}^\theta)^n.$$

Taking into account the facts that $\nabla c(\theta_\star)/c(\theta_\star) = y_{\mathrm{obs}}$ and $\nabla f_\theta(y) = y\mathrm{e}^{\theta y}$ we obtain that (2.11) becomes $h_\star(y) \propto |y - y_{\mathrm{obs}}|\mathrm{e}^{\theta y}$ (factor D^{-1} is a scalar so can be omitted). In other words, the optimum instrumental distribution for AIS MCML, expressed in terms of $p = \mathrm{e}^\theta/(1+\mathrm{e}^\theta)$ is

$$\mathbb{P}_{Y \sim h_\star}(Y = y) \propto \binom{n}{y} |y - y_{\mathrm{obs}}| p^y (1-p)^{n-y}.$$

3 Generalized Adaptive Scheme

Importance sampling, even in its adaptive version (AIS), suffers from the degeneracy of weights. To compute the importance weights $f_\theta(Y_m)/h_{\psi_m}(Y_m)$ we have to know norming constants for every h_{ψ_m} (or at least their ratios). This requirement severly restricts our choice of the family of instrumental densities h_ψ. Available instrumental densities are far from $h_\star$ and far from $f_\theta/c(\theta)$. Consequently the weights tend to degenerate (most of them are practically zero, while a few are very large). This effectively makes AIS in its basic form impractical. To obtain practically applicable algorithms, we can generalize AIS as follows. In the same situation as in Sect. 2, instead of the AIS estimate given by (2.3), we consider a more general Monte Carlo estimate of $c(\theta)$ of the form

$$\hat{c}_m(\theta) = \frac{1}{m} \sum_{j=1}^{m} \hat{d}(\theta, \psi_j), \tag{3.1}$$

where the summands $\hat{d}(\theta, \psi_j)$ are computed by an MC method to be specified later. For now let us just assume that this method depends on a control parameter ψ which may change at each step. A general adaptive algorithm is the following:

```
Algorithm AdapMCML

1. Set an initial value of $\psi_1, m = 1, \hat{c}_0(\theta) \equiv 0$.
2. Compute an "incremental estimate" $\hat{d}(\theta, \psi_m)$.
3. Update the approximation of $\hat{c}_m(\theta)$:

       $\hat{c}_m(\theta) = \frac{m-1}{m}\hat{c}_{m-1}(\theta) + \frac{1}{m}\hat{d}(\theta, \psi_m)$.

4. Update $\psi$: choose $\psi_{m+1}$ based on the history of the
   simulation.
5. $m = m + 1$; go to 2.
```

AdapIS in Sect. 2 is a special case of AdapMCML which is obtained by letting $\hat{d}(\theta, \psi_m) = f_\theta(Y_m)/h_{\psi_m}(Y_m)$.

3.1 Variance Reduction Via Resampling and MCMC

The key property of the AIS exploited in Sect. 2 is the martingale structure implied by (2.4) and (2.6). The main asymptotic results generalize if *given* ψ, *the estimates of* $c(\theta)$ *and its derivatives are conditionally unbiased*. We propose an algorithm for computing $\hat{d}$ in (3.1) which has the unbiasedness property and is more efficient than simple AIS. To some extent it is a remedy for the problem of weight degeneracy and reduces the variance of Monte Carlo approximations. As before, consider a family of "instrumental densities" h_ψ. Assume they are properly normalized ($\int h_\psi = 1$) and the control parameter ψ belongs the same space as the parameter of interest θ ($\Psi = \Theta$). Further assume that for every ψ we have at our disposal a Markov kernel P_ψ on $\mathcal{Y}$ which preserves distribution $\pi_\psi = f_\psi/c(\psi)$, i.e. $f_\psi(y)\mathrm{d}y = \int f_\psi(y')P_\psi(y', \mathrm{d}y)\mathrm{d}y'$. Let us fix ψ. This is a setup in which we can apply the following importance sampling-resampling algorithm ISReMC:

```
                         Algorithm ISReMC
1. Sample $Y_1, \dots, Y_l \sim h_\psi$.
2. Compute the importance weights $W_i = w(Y_i) = \dfrac{f_\psi(Y_i)}{h_\psi(Y_i)}$ and put
   $W_\bullet = \sum_{i=1}^{l} W_i$.
3. Sample $Y_1^\star, \dots, Y_r^\star \sim \sum_{i=1}^{l} \delta_{Y_i}(\cdot) W_i / W_\bullet$ [Discrete distribution
   with mass $W_i / W_\bullet$ at point $Y_i$].
4. For $k = 1, \dots, r$ generate a Markov chain trajectory,
   starting from $Y_k^\star$ and using kernel $P_\psi$:
       $Y_k^\star = Y_k^0, Y_k^1, \dots, Y_k^s, Y_k^{s+1}, \dots, Y_k^{s+n}$.
   Compute $\hat{d}(\theta, \psi)$ given by
       $\hat{d}(\theta, \psi) = \frac{W_\bullet}{l} \frac{1}{r} \sum_{k=1}^{r} \frac{1}{n} \sum_{u=s+1}^{s+n} \frac{f_\theta(Y_k^u)}{f_\psi(Y_k^u)}$.   (3.2)
```

This algorithm combines the idea of resampling (borrowed from sequential MC; steps 2 and 3) with computing ergodic averages in multistart MCMC (step 4; notice that s is a burn-in and n is the actual used sample size for a single MCMC run, repeated r times). More details about ISReMC are in [7]. In our context it is sufficient to note the following key property of this algorithm.

Lemma 1 *If $\hat{d}(\theta, \psi)$ is the output of IReMC then for every θ and every ψ,*

$$\mathbb{E}\hat{d}(\theta, \psi) = c(\theta).$$

If Assumption 3 holds then also

$$\mathbb{E}\nabla\hat{d}(\theta, \psi) = \nabla c(\theta), \qquad \mathbb{E}\nabla^2\hat{d}(\theta, \psi) = \nabla^2 c(\theta).$$

Proof We can express function $c(\theta)$ and its derivatives as "unnormalized expectations" with respect to the probability distribution with density $\pi_\psi = f_\psi / c(\psi)$:

$$c(\theta) = \mathbb{E}_{Y \sim \pi_\psi} \frac{f_\theta(Y)}{f_\psi(Y)} c(\psi),$$

$$\nabla c(\theta) = \mathbb{E}_{Y \sim \pi_\psi} \frac{\nabla f_\theta(Y)}{f_\psi(Y)} c(\psi), \qquad \nabla^2 c(\theta) = \mathbb{E}_{Y \sim \pi_\psi} \frac{\nabla^2 f_\theta(Y)}{f_\psi(Y)} c(\psi).$$

Let us focus on $\mathbb{E}\hat{d}(\theta, \psi)$. Write

$$a(y) = \mathbb{E}\left(\frac{1}{n} \sum_{u=s+1}^{s+n} \frac{f_\theta(Y^u)}{f_\psi(Y^u)} \middle| Y^0 = y \right) \tag{3.3}$$

for the expectation of a *single* MCMC estimate started at $Y^0 = y$. Kernel P_ψ preserves π_ψ by assumption, therefore $\mathbb{E}_{Y\sim\pi_\psi} a(Y) = \mathbb{E}_{Y\sim\pi_\psi} f_\theta(Y)/f_\psi(Y) = c(\theta)/c(\psi)$. Put differently, $\int a(y) f_\psi(y) \mathrm{d}y = c(\theta)$.

We make a simple observation that

$$\mathbb{E}\left(\hat{d}(\theta, \psi) \middle| Y_1, \ldots, Y_l, Y_1^\star, \ldots, Y_r^\star\right) = \frac{W_\bullet}{l} \frac{1}{r} \sum_{k=1}^{r} a(Y_k^\star).$$

This conditional expectation takes into account only randomness of the MCMC estimate in step 4 of the algorithm. Now we consecutively "drop the conditions":

$$\mathbb{E}\left(\hat{d}(\theta, \psi) \middle| Y_1, \ldots, Y_l\right) = \frac{W_\bullet}{l} \sum_{i=1}^{l} a(Y_i) \frac{W_i}{W_\bullet} = \frac{1}{l} \sum_{i=1}^{l} a(Y_i) W_i.$$

The expectation above takes into account the randomness of the resampling in step 3. Finally, since $Y_i \sim h_\psi$ in step 1, we have

$$\begin{aligned}\mathbb{E}\hat{d}(\theta, \psi) &= \mathbb{E}a(Y_i)W_i = \mathbb{E}_{Y\sim h_\psi} a(Y) \frac{f_\psi(Y)}{h_\psi(Y)} \\ &= \int a(y) f_\psi(y) \mathrm{d}y = c(\theta).\end{aligned}$$

This ends the proof for $\hat{d}$. Exactly the same argument applies to $\nabla \hat{d}$ and $\nabla^2 \hat{d}$.

We can embed the unbiased estimators produced by ISReMC in our general adaptive scheme AdapMCML. At each step m of the adaptive algorithm, we have a new control parameter ψ_m. We generate a sample from h_{ψ_m}, compute weights, resample and run MCMC using ψ_m. Note that the whole sampling scheme at stage m (including computation of weights) depends on ψ_m but not on θ. In the adaptive algorithm random variable ψ_{m+1} is $\mathcal{F}_m$ measurable, where $\mathcal{F}_m$ is the history of simulation up to stage m. Therefore the sequence of incremental estimates $\hat{d}(\theta, \psi_m)$ satisfies, for every $\theta \in \Theta$,

$$\mathbb{E}(\hat{d}(\theta, \psi_{m+1}) | \mathcal{F}_m) = c(\theta). \tag{3.4}$$

Moreover, first and second derivatives exist and

$$\mathbb{E}(\nabla\hat{d}(\theta,\psi_{m+1})|\mathcal{F}_m)=\nabla c(\theta),\qquad \mathbb{E}(\nabla^2\hat{d}(\theta,\psi_{m+1})|\mathcal{F}_m)=\nabla^2 c(\theta). \tag{3.5}$$

Formulas (3.4) and (3.5) are analogues of (2.4) and (2.6).

3.2 Asymptotics of Adaptive MCML

In this subsection we restrict our considerations to *exponential families on finite spaces*. This will allow us to prove main results without formulating complicated technical assumptions (integrability conditions analoguous to Assumption 8 would be cumbersome and difficult to verify). Some models with important applications, such as *autologistic* one, satisfy the assumptions below.

Assumption 10 Let

$$f_\theta(y)=\exp[\theta^\top t(y)],$$

where $t(y)\in\mathbb{R}^d$ is the vector of sufficient statistics and $\theta\in\Theta=\mathbb{R}^d$. Assume that y belongs to a finite space $\mathcal{Y}$.

Now, since $\mathcal{Y}$ is finite (although possibly very large),

$$c(\theta)=\sum_y \exp[\theta^\top t(y)].$$

Note that Assumption 3 is automatically satisfied.

Assumption 11 Control parameters ψ belong to a compact set $\Psi\subset\mathbb{R}^d$.

We consider algorithm `AdapMCML` with incremental estimates $\hat{d}$ produced by `ISReMC`. The likelihood ratio in (3.2) and its derivatives assume the following form:

$$\begin{aligned}\frac{f_\theta(Y)}{f_\psi(Y)}&=\exp[(\theta-\psi)^\top t(Y)],\\ \frac{\nabla f_\theta(Y)}{f_\psi(Y)}&=t(Y)\exp[(\theta-\psi)^\top t(Y)],\\ \frac{\nabla^2 f_\theta(Y)}{f_\psi(Y)}&=t(Y)t(Y)^\top\exp[(\theta-\psi)^\top t(Y)]\end{aligned} \tag{3.6}$$

(the derivatives are with respect to θ, with ψ fixed). Assumptions 11 and 10 together with Assumption 6 imply that $\hat{d}(\theta,\psi_j)$ are uniformly bounded, if θ belongs to a compact set. Indeed, the importance weights W_i in (3.2) are uniformly bounded by Assumptions 6 and 11. Formula (3.6) shows that the ratios $f_\theta(y)/f_{\psi_j}(y) = \exp[(\theta - \psi_j)^\top t(y)]$ are also uniformly bounded for ψ_j and θ belonging to bounded sets. Since the statistics $t(y)$ are bounded, the same argument shows that $\nabla\hat{d}(\theta, \psi_j)$ and $\nabla^2\hat{d}(\theta, \psi_j)$ are uniformly bounded, too.

For exponential families, $\log c(\theta)$ and $\log \hat{c}_m(\theta)$ are convex functions. It is a well known property of exponential family that $\nabla^2 \log c(\theta) = \mathrm{VAR}_{Y\sim\pi_\theta} t(Y)$ and thus it is a nonnegative definite matrix. A closer look at $\hat{c}_m(\theta)$ reveals that $\nabla^2 \log \hat{c}_m(\theta)$ is also a variance-covariance matrix with respect to some discrete distribution. Indeed, it is enough to note that $\hat{c}_m(\theta)$ is of the form

$$\hat{c}_m(\theta) = \sum_{j,k,u} \exp[\theta^\top t_{j,k,u}] a_{j,k,u},$$

for some $t_{j,k,u} \in \mathbb{R}^d$ and $a_{j,k,u} > 0$ (although if `ISReMC` within `AdapMCML` is used to produce $\hat{c}_m(\theta)$ then $t_{j,k,u}$ and $a_{j,k,u}$ are quite complicated random variables depending on ψ_j).

Let $\hat{\theta}_m$ be a maximizer of $\ell_m(\theta) = \theta^\top t(y_{\mathrm{obs}}) - \log \hat{c}_m(\theta)$ and assume that $\theta_\star$ is the unique maximizer of $\ell(\theta) = \theta^\top t(y_{\mathrm{obs}}) - \log c(\theta)$.

Proposition 3 *If Assumptions 6, 10 and 11 hold, then $\hat{\theta}_m \to \theta_\star$ almost surely.*

Proof Boundedness of $\hat{d}(\theta, \psi_m)$ for a fixed θ together with (3.4) implies that $\hat{d}(\theta, \psi_m) - c(\theta)$ is a bounded sequence of martingale differences. It satisfies the assumptions of SLLN for martingales in Appendix A. Therefore $\hat{c}_m(\theta) \to c(\theta)$. Consequently, we also have $\ell_m(\theta) \to \ell(\theta)$, pointwise. Pointwise convergence of convex functions implies uniform convergence on compact sets [13, Theorem 10.8]. The conclusion follows immediately.

Theorem 4 *If Assumptions 5–7, 10 and 11 hold, then*

$$\sqrt{m}(\hat{\theta}_m - \theta_\star) \to \mathcal{N}(0, D^{-1} V D^{-1}) \text{ in distribution},$$

where $D = \nabla^2\ell(\theta_\star)$ and

$$V = \frac{1}{c(\theta_\star)^2} \mathrm{VAR}\left[\nabla\hat{d}(\theta_\star, \psi_\star) - \frac{\nabla c(\theta_\star)}{c(\theta_\star)}\hat{d}(\theta_\star, \psi_\star)\right],$$

where $\hat{d}(\theta_\star, \psi_\star)$ is a result of the IS/Resampling algorithm `ISReMC`, described in the previous subsection, with $\psi = \psi_\star$ and $\theta = \theta_\star$.

Note that $\hat{d}(\theta_\star, \psi_\star)$ is a purely imaginary object, being a result of an algorithm initialized at a “limiting instrumental parameter” $\psi_\star$ and evaluated at the “true MLE” $\theta_\star$, both unknown. It is introduced only to concisely describe the variance/covariance matrix V. Note also that $\nabla c(\theta_\star)/c(\theta_\star)$ is equal to $t(y_{\mathrm{obs}})$, the *observed* value of the sufficient statistic.

Proof (of Theorem 4) The proof is similar to that of Theorem 3, so we will not repeat all the details. The key argument is again based on SLLN and CLT for martingales (see Appendix A). In the present situation we have more complicated estimators $\hat{d}(\theta, \psi_j)$ than in Theorem 3. They are now given by (3.2). On the other hand, we work under the assumption that f_θ is an exponential family on a finite state space $\mathcal{Y}$. This implies that conditions (3.4) and (3.5) are fulfilled and the martingale differences therein are uniformly bounded (for any fixed θ and also for θ running through a compact set). Concavity of $\ell_m(\theta)$ and $\ell(\theta)$ further simplifies the argumentation.

As in the proof of Theorem 3, we claim that (2.8) and (2.9) hold. The first of these conditions, (2.8), is justified exactly in the same way: by applying the CLT to the numerator and SLLN to the denominator of (2.10). Now, we consider martingale differences given by

$$\xi_j = \nabla \hat{d}(\theta_\star, \psi_j) - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \hat{d}(\theta_\star, \psi_j).$$

It follows from the discussion preceding the theorem that ξ_j are uniformly bounded, so the CLT can be applied. Similarly, SLLN can be applied to $\hat{d}(\theta_\star, \psi_j) - c(\theta_\star)$.

Assumption (9) holds because third order derivatives of $\log \hat{c}_m(\theta)$ are uniformly bounded in the neighbourhood of $\theta_\star$. This allows us to infer condition (2.9) in the same way as in the proof of Theorem 3.

Note also that we do not need Assumption 4. To deduce the conclusion of the theorem from (2.9) we have to know that $\hat{\theta}_m$ is square-root consistent. But this follows from the facts that $\ell_m(\theta)$ is concave and the maximizer of the quadratic function $-(\theta-\theta_\star)^\top \nabla \ell_m(\theta_\star) - \frac{1}{2}(\theta-\theta_\star)^\top D(\theta-\theta_\star)$ in (2.9) is square-root consistent by (2.8).

4 Simulation Results

In a series of small scale simulation experiments, we compare two algorithms. The first one, used as a "Benchmark" is a non-adaptive MCML. The other is `AdapMCML` which uses `ISReMC` estimators, as described in Sect. 3. Synthetic data used in our study are generated from autologistic model, described below. Both algorithms use Gibbs Sampler (GS) as an MCMC subroutine and both use Newton-Raphson iterations to maximize MC log-likelihood approximations.

4.1 Non-adaptive and Adaptive Newton-Raphson-Type Algorithms

Well-known Newton-Raphson (NR) method in our context updates points θ_m approximating maximum of the log-likelihood as follows:

$$\theta_{m+1} = \theta_m + \nabla^2 \ell_m(\theta_m)^{-1} \nabla \ell_m(\theta_m),$$

where ℓ_m is given by (2.1).

Non-adaptive algorithms are obtained when some fixed value of the "instrumental parameter" is used to produce MC samples. Below we recall a basic version of such an algorithm, proposed be Geyer [4] and examined e.g. in [9]. If we consider an exponenial family given by Assumption 10, then $\ell_m(\theta) = \theta^\top t(y_{\text{obs}}) - \log \hat{c}_m(\theta)$. Let ψ be fixed and $Y_0, Y_1, \dots, Y_s, \dots, Y_{s+m}$ be samples approximately drawn from distribution $\pi_\psi \propto f_\psi$. In practice an MCMC method is applied to produce such samples, s stands for a burn-in. In all our experiments the MCMC method is a deterministic scan Gibbs Sampler (GS). Now, we let

$$\hat{c}_m(\theta) \propto \frac{1}{m} \sum_{u=s+1}^{s+m} \exp[(\theta - \psi)^\top t(Y_u)].$$

Consequently, if $\omega_u(\theta) = \exp[(\theta - \psi)^\top t(Y_u)]$ and $\omega_\bullet(\theta) = \sum_{u=s+1}^{s+m} \omega_u(\theta)$, then the derivatives of the log-likelihood are expressed via weighted moments,

$$\nabla \ell_m(\theta) = t(y_{\text{obs}}) - \overline{t(Y)}, \quad \overline{t(Y)} = \frac{1}{\omega_\bullet(\theta)} \sum_{u=s+1}^{s+m} \omega_u(\theta) t(Y_u),$$

$$\nabla^2 \ell_m(\theta) = -\frac{1}{\omega_\bullet(\theta)} \sum_{u=s+1}^{s+m} \omega_u(\theta)(t(Y_u - \overline{t(Y)})(t(Y_u) - \overline{t(Y)})^\top.$$

The adaptive algorithm uses $\hat{c}_m(\theta)$ given by (3.1), with summands $\hat{d}(\theta, \psi_j)$ computed by `ISReMC`, exactly as described in Sect. 3. The MCMC method imbedded in `ISReMC` is GS, the same as in the non-adaptive algorithm. Importance sampling distribution h_ψ in steps 1 and 2 of `ISReMC` is pseudo-likelihood, described by formula (4.1) in the next subsection. Computation of ψ_{m+1} in step 4 of `AdapMCML` uses one NR iteration: $\psi_{m+1} = \psi_m + \nabla^2 \ell_m(\psi_m)^{-1} \nabla \ell_m(\psi_m)$, where ℓ_m is given by (2.1) with $\hat{c}_m$ produced by `AdapMCML`.

4.2 *Methodology of Simulations*

For our experiments we have chosen the autologistic model, one of chief motivating examples for MCML. It is given by a probability distribution on $\mathcal{Y} = \{0, 1\}^{d \times d}$ proportional to

$$f_\theta(y) = \exp\left(\theta_0 \sum_r y^{(r)} + \theta_1 \sum_{r \sim s} y^{(r)} y^{(s)} \right),$$

Table 1 Sufficient statistics, maximum log-likelihood and MPL estimate for example with $d = 10$

Statistic T	ML $\theta_\star$	Log-Lik $\ell(\theta_\star)$	MPL $\hat{\theta}$
$(59, 74)$	$(-1.21, 0.75)$	-15.889991	$(-1.07, 0.66)$

Table 2 Sufficient statistics, maximum log-likelihood and MPL estimate for example with $d = 15$

Statistic T	ML $\theta_\star$	Log-Lik $\ell(\theta_\star)$	MPL $\hat{\theta}$
$(142, 180)$	$(-0.46, 0.43)$	12.080011	$(-0.57, 0.54)$

where $r \sim s$ means that two points r and s in the $d \times d$ lattice are neighbours. The pseudo-likelihood h_ψ is given by

$$h_\psi(y) \propto \prod_r \exp\left(\theta_0 y^{(r)} + \theta_1 \sum_{s:r\sim s} y^{(r)} y^{(s)}_{\text{obs}}\right). \tag{4.1}$$

In our study we considered lattices of dimension $d = 10$ and $d = 15$. The values of sufficient statistics $T = \left(\sum_r y^{(r)}, \sum_{r\sim s} y^{(r)} y^{(s)}\right)$, exact ML estimators $\theta_\star$ and maxima of the log-likelihoods are in the Tables 1 and 2. We report results of several repeated runs of a "benchmark" non-adaptive algorithm and our adaptive algorithm. The initial points are (1) the maximum pseudo-likelihood (MPL) estimate, denoted by $\hat{\theta}$ (also included in the tables) and (2) point $(0, 0)$. Number of runs is 100 for $d = 10$ and 25 for $d = 15$. Below we describe the parameters and results of these simulations. Note that we have chosen parameters for both algorithms in such a way which allows for a "fair comparison", that is the amount of computations and number of required samples are similar for the benchmark and adaptive algorithms.

For $d = 10$: In benchmark MCML, we used 1000 burn-in and 39,000 collected realisations of the Gibbs sampler; then 20 iterations of Newton-Raphson were applied. `AdapMCML` had 20 iterations; parameters within `ISReMC` were $l = 1000$, $r = 1$, $s = 100$, $n = 900$.

The results are shown in Figs. 1 and 2.

For $d = 15$: In benchmark MCML, we used 10,000 burn-in and 290,000 collected realisations of the Gibbs sampler; then 10 iterations of Newton-Raphson were applied. `AdapMCML` had 10 iterations; parameters within `ISReMC` were $l =$10,000, $r = 1$, $s = 1000$, $n =$19,000.

The results are shown in Figs. 3 and 4. The benchmark algorithm started from 0 for $d = 15$ failed, so only the results for the adaptive algorithm are given in the right parts of Figs. 3 and 4.

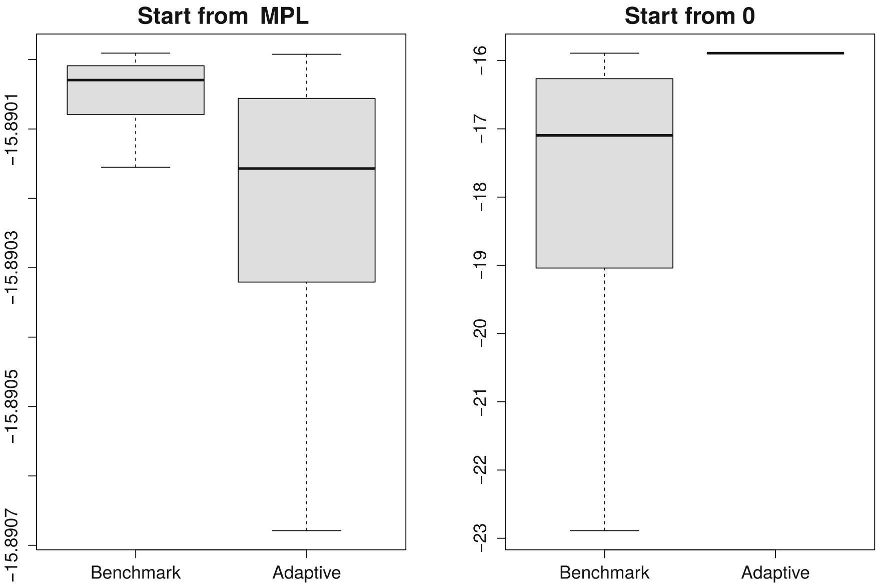


Fig. 1 Log-likelihood at the output of MCML algorithms; $d = 10$; 100 runs

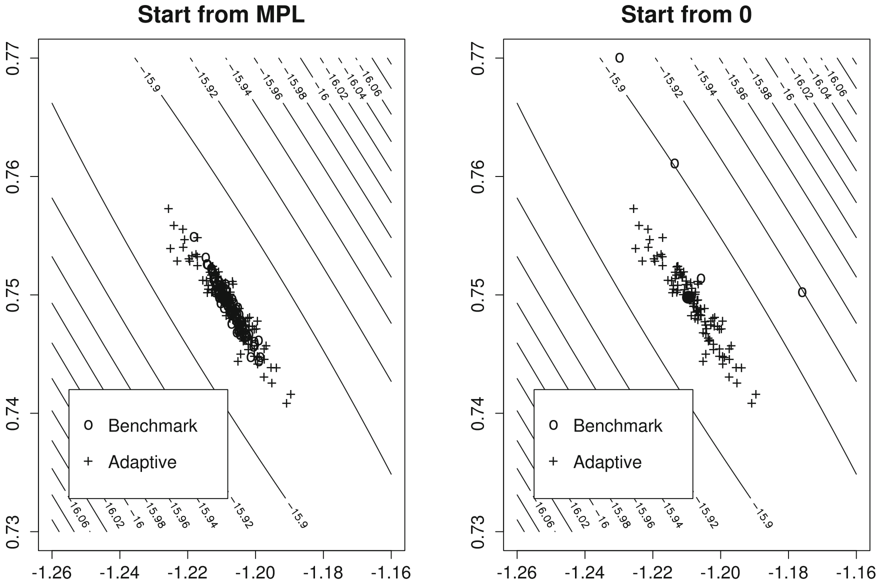


Fig. 2 Output of MCML algorithms; $d = 10$; 100 repetitions

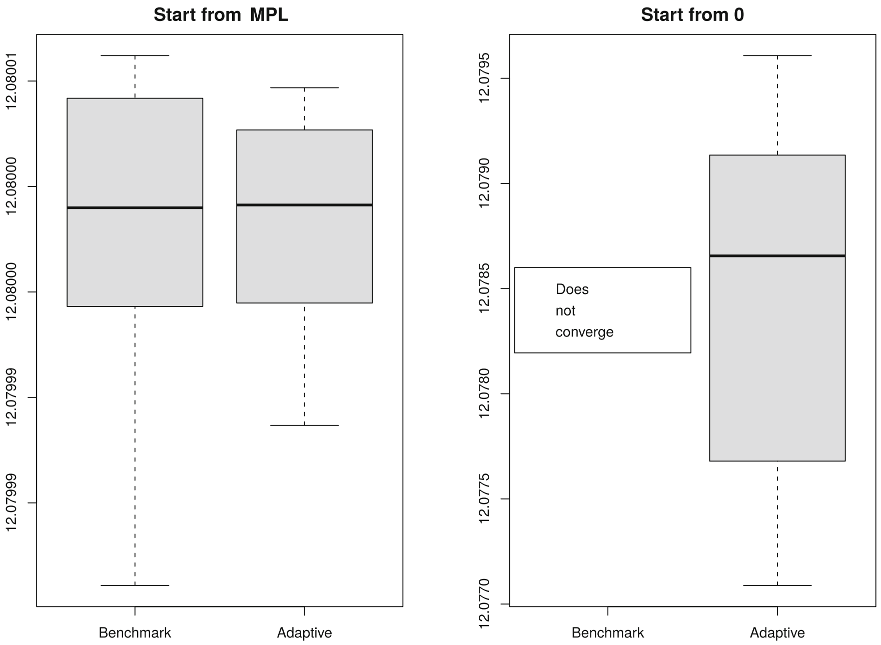


Fig. 3 Log-likelihood at the output of MCML algorithms; $d = 15$; 25 repetitions

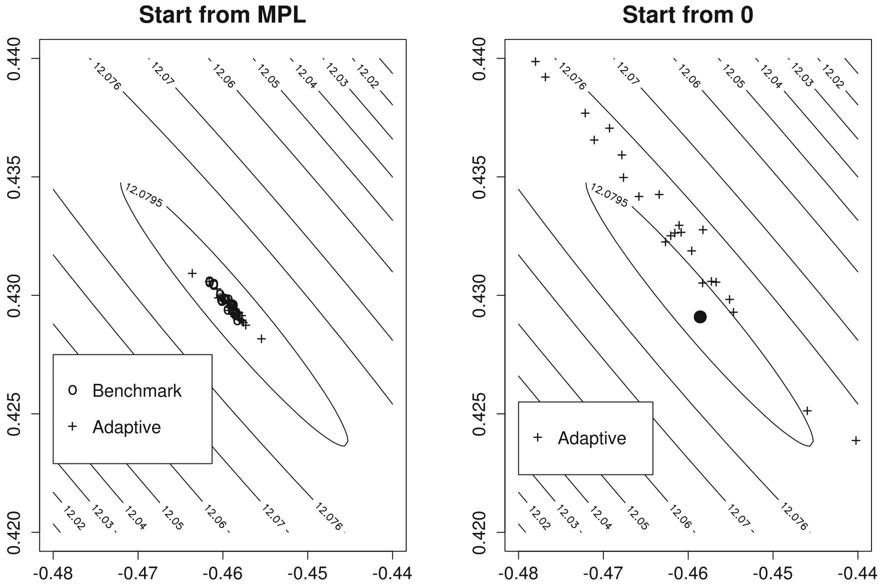


Fig. 4 Output of MCML algorithms; $d = 15$; 25 repetitions

4.3 Conclusions

The results of our simulations allow to draw only some preliminary conclusions, because the range of experiments was limited. However, some general conclusions can be rather safely formulated. The performance of the benchmark, non-adaptive algorithm crucially depends on the choice of starting point. It yields quite satisfactory results, if started sufficiently close tho the maximum likelihood, for example from the maximum pseudo-likelihood estimate. Our adaptive algorithm is much more robust and stable in this respect. If started from a good initial point, it may give slightly worse results than the benchmark, but still is satisfactory (see Fig. 2). However, when the maximum pseudo-likelihood estimate is not that close to the maximum likelihood point, the adaptive algorithm yields an estimate with a lower variance (see Fig. 2). When started at a point distant from the maximum likelihood, such as 0, it works much better than a non-adaptive algorithm. Thus the algorithm proposed in our paper can be considered as more universal and robust alternative to a standard MCML estimator.

Finally let us remark that there are several possibilities of improving our adaptive algorithm. Some heuristically justified modifications seem to converge faster and be more stable than the basic version which we described. Modifications can exploit the idea of resampling in a different way and reweigh past samples in subsequent steps. Algorithms based on stochastic approximation, for example such as that proposed in [16], can probably be improved by using Newton-Raphson method instead of simple gradient descent. However, theoretical analysis of such modified algorithms becomes more difficult and rigorous theorems about them are not available yet. This is why we decided not to include these modified algorithms in this paper. Further research is needed to bridge a gap between practice and theory of MCML.

Acknowledgments This work was partially supported by Polish National Science Center No. N N201 608 740.

Appendix A: Martingale Limit Theorems

For completeness, we cite the following martingale central limit theorem (CLT):

Theorem A.1 ([8, Theorem 2.5]) *Let $X_n = \xi_1 + \cdots + \xi_n$ be a mean-zero (vector valued) martingale. If there exists a symmetric positive definite matrix V such that*

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\Big(\xi_j\xi_j^T|\mathcal{F}_{j-1}\Big) \xrightarrow{\text{p}} V, \tag{A.1}$$

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\Big(\xi_j\xi_j^T\mathbf{1}_{|\xi_j|>\varepsilon\sqrt{n}}\,|\mathcal{F}_{j-1}\Big)\xrightarrow{\text{p}}0 \quad \textit{for each } \varepsilon>0, \tag{A.2}$$

then

$$\frac{X_n}{\sqrt{n}}\xrightarrow{\text{d}}\mathcal{N}(0,V).$$

The Lindeberg condition (A.2) can be replaced by a stronger Lyapunov condition

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\Big(|\xi_j|^{2+\alpha}|\mathcal{F}_{j-1}\Big)\le M \quad \text{for some } \alpha>0 \text{ and } M<\infty. \tag{A.3}$$

A simple consequence of [6, Theorem 2.18] (see also [3]) is the following strong law of large numbers (SLLN).

Theorem A.2 *Let $X_n=\xi_1+\cdots+\xi_n$ be a mean-zero martingale. If*

$$\sup_j\mathbb{E}\Big(|\xi_j|^{1+\alpha}|\mathcal{F}_{j-1}\Big)\le M \quad \textit{for some } \alpha>0 \textit{ and } M<\infty$$

then

$$\frac{X_n}{n}\xrightarrow{\text{a.s.}}0.$$

References

1. Attouch H (1984) Variational convergence of functions and operators. Pitman, Boston
2. Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc B 36:192–236
3. Chow YS (1967) On a strong law of large numbers for martingales. Ann Math Stat 38:610
4. Geyer CJ (1994) On the convergence of Monte Carlo maximum likelihood calculations. J R Stat Soc B 56:261–274
5. Geyer CJ, Thompson EA (1992) Constrained Monte Carlo maximum likelihood for dependent data. J R Stat Soc B 54:657–699
6. Hall P, Heyde CC (1980) Martinagale limit theory and its application. Academic Press, New York
7. Miasojedow B, Niemiro W (2014) Debiasing MCMC via importance sampling-resampling. In preparation
8. Helland IS (1982) Central limit theorems for martingales with discrete or continuous time. Scand J Stat 9:79–94
9. Huffer FW, Wu H (1998) Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. Biometrics 54:509–524
10. Møller BJ, Pettitt AN, Reeves R, Berthelsen KK (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika 93:451–458
11. Pettitt AN, Weir IS, Hart AG (2002) A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. Stat Comput 12:353–367

12. Pollard D (1984) Convergence of stochastic processes. Springer, New York
13. Rockafellar RT (1970) Convex analysis. Princeton University Press, Princeton
14. Rockafellar TJ, Wets RJ-B (2009) Variational analysis, 3rd edn. Springer, New York
15. Wu H, Huffer FW (1997) Modeling the distribution of plant species using the autologistic regression model. Environ Ecol Stat 4:49–64
16. Younes L (1988) Estimation and annealing for Gibbsian fields. Annales de l'I H P sec B 24(2):269–294
17. Zalewska M, Niemiro W, Samoliński B (2010) MCMC imputation in autologistic model. Monte Carlo Methods Appl 16:421–438

What Do We Choose When We Err? Model Selection and Testing for Misspecified Logistic Regression Revisited

Jan Mielniczuk and Paweł Teisseyre

Abstract The problem of fitting logistic regression to binary model allowing for missppecification of the response function is reconsidered. We introduce two-stage procedure which consists first in ordering predictors with respect to deviances of the models with the predictor in question omitted and then choosing the minimizer of Generalized Information Criterion in the resulting nested family of models. This allows for large number of potential predictors to be considered in contrast to an exhaustive method. We prove that the procedure consistently chooses model t^* which is the closest in the averaged Kullback-Leibler sense to the true binary model t. We then consider interplay between t and t^* and prove that for monotone response function when there is genuine dependence of response on predictors, t^* is necessarily nonempty. This implies consistency of a deviance test of significance under misspecification. For a class of distributions of predictors, including normal family, Rudd's result asserts that $t^* = t$. Numerical experiments reveal that for normally distributed predictors probability of correct selection and power of deviance test depend monotonically on Rudd's proportionality constant η.

Keywords Incorrect model specification · Variable selection · Logistic regression

J. Mielniczuk (✉)
Faculty of Mathematics and Information Science, Warsaw University of Technology,
Koszykowa 75, 00-662 Warsaw, Poland
e-mail: miel@ipipan.waw.pl

J. Mielniczuk · P. Teisseyre
Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5,
01-248 Warsaw, Poland
e-mail: teisseyrep@ipipan.waw.pl
url: http://www.ipipan.waw.pl

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_15

1 Introduction

We consider a general binary regression model in which responses $y \in \{0, 1\}$ are related to explanatory variables $\mathbf{x} = (1, x_1, \ldots, x_p)' \in R^{p+1}$ by the equation

$$P(y = 1|\mathbf{x}) = q(\mathbf{x}'\boldsymbol{\beta}), \tag{1}$$

where vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is an unknown vector of parameters and $q : R \to (0, 1)$ is a certain unknown response function. To the data pertaining to (1) we fit the logistic regression model i.e. we postulate that the posterior probability that $y = 1$ given $\mathbf{x}$ is of the form

$$p(\mathbf{x}'\boldsymbol{\gamma}) = \exp(\mathbf{x}'\boldsymbol{\gamma})/[1 + \exp(\mathbf{x}'\boldsymbol{\gamma})], \tag{2}$$

where $\boldsymbol{\gamma} \in R^{p+1}$ is a parameter. Our main interest here is the situation when the logistic model is misspecified i.e. $p \neq q$. Let $t = \{0\} \cup \{1 \leq k \leq p : \beta_k \neq 0\}$ be the true model i.e. consisting of indices of nonzero coefficients corresponding to true predictors and of the intercept denoted by 0. Our task may be either to identify model t when incorrectly specified model (2) is fitted or, less ambitiously, to verify whether t contains indices corresponding to predictors i.e. whether response depends on predictors at all. The situation of incorrect model specification is of importance because of obvious reasons as in real applications usually we have no prior knowledge about data generation process and, moreover, goodness-of-fit checks may yield inconclusive results. Thus investigating to what extent selection and testing procedures are resistant to response function misspecification is of interest. This is especially relevant with large number of possible features and sparsity when selecting true predictors is a challenge in itself and is further exacerbated by possible model misspecification. Moreover, some data generation mechanisms lead directly to misspecified logistic model. As an example we mention [6] who consider the case of logistic model when each response is mislabeled with a certain fixed probability.

In the paper we consider selection procedures specially designed for large p scenario which use Generalized Information Criterion (GIC). This criterion encompasses, for specific choices of parameters, such widely used criteria as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC is known to overestimate the dimension of the true model (see e.g. [4]) whereas BIC in the case of correctly specified linear model with fixed p is consistent [7]. There are many modifications of AIC and BIC which among others are motivated by the phenomenon that for large p depending on the sample size BIC also choses too large number of variables. We mention in particular modified BIC [3, 23], Extended BIC (EBIC) which consists in adding a term proportional to log p to BIC [8, 9] and Risk Inflation Criterion [15]. Qian and Field [20] consider GIC and proved its consistency under correct specification. In this line of research [9] propose minimization of EBIC over all possible subsets variables of sizes not larger than k when k is some sufficiently large number. However, this approach becomes computationally prohibitive for even moderate k. Other important approach is based on l_1-penalized loglikelihood and its extensions and modifications such as Elastic Net (see [24]) and SCAD [14]. It is known that l_1-penalization leads to cancelation of some coefficients and thus can be considered as model selection method. For discussion of other approaches we refer to [5, 10, 17] and references there.

The aims of the paper are twofold. We first introduce two-step modification of a procedure based on GIC, the minimizer of which over the family of all possible models is used as a selector of relevant variables. In the case when number of possible predictors is large such an approach is practically unfeasible due to high computational cost of calculating GIC for all possible subsets. This is a reason, likely the only one, why these methods are not frequently used and sequential greedy methods are applied in practice. However, greedy methods lack theoretical underpinning and it is known that they may miss true predictors. We thus propose a specific two-stage greedy method which consists in first ranking the predictors according to residual deviances of the models containing all variables but the considered one. Then in the second stage GIC is minimized over the nested family of models pertaining to increasing sets of the most important variables. We prove that such procedure picks with probability tending to 1 the logistic model t^* which minimizes averaged Kullback-Leibler distance from the binary model (1). This is to the best of our knowledge the first formal result on the consistency of greedy selection procedure for logistic regression even in the case when $p = q$. As a by-product we obtain the known result concerning behaviour of GIC optimized over the family of all models due to [22]. As in their paper the very general framework is considered for which stringent assumptions are needed we note that it is possible to prove the result under much weaker conditions (cf. their Proposition 4.2 (i), (ii) and Theorem 2 below). In view of the result the nature of the interplay between t^* and t becomes relevant. However, it seems that the problem, despite its importance, has failed to attract much attention. Addressing this question, admittedly partially, is the second aim of the paper. We discuss Rudd's (1983) result in this context which states that for certain distributions of predictors $\boldsymbol{\beta}^* = \eta\boldsymbol{\beta}$ for some $\eta \in R$, where $\boldsymbol{\beta}^*$ which minimizes averaged Kullback-Leibler distance from the binary model to logistic regressions. This obviously implies that $t^* = t$ if $\eta \neq 0$. As our main result in this direction we prove in Theorem 4 if t contains genuine regressors so does t^* provided that q is monotone and not constant. This implies in particular that in such a case significance test for regressors constructed under logistic model is consistent under misspecification. We also discuss the relevance of proved results in practice by investigating probability of correct model selection for two-stage procedure and power of test of significance for moderate sample sizes. In particular, we empirically verify that, surprisingly, misspecification of the model may lead to larger probabilities of correct selection and positive selection rate than for correct specification and stress the importance of the proportionality constant η in this context. Namely, it turns out that this phenomenon occurs mostly in the cases when $\eta > 1$. Moreover, we established that probability of correct selection and power of deviance test depend monotonically on η.

Generalization to the case when p is large in comparison to n is left for further study. As the fitting of the full model in the first stage of the procedure excludes its application when $p > n$ an initial screening of variables which is commonly done in applications (see e.g. [9]) would be necessary.

The paper is structured as follows. Section 2 contains preliminaries, in Sect. 3 we introduce and prove consistency of two-step greedy GIC procedure. Interplay between t and t^* is discussed in Sect. 4 together with its consequence for consistency of deviance test under misspecification. In Sect. 5 we describe our numerical experiments and Appendix contains proofs of auxiliary lemmas.

2 Preliminaries

Observe that the first coordinate of $\boldsymbol{\beta}$ in (1) corresponds to the intercept and remaining coefficients to genuine predictors which are assumed to be random variables. We assume that $\boldsymbol{\beta}$ is uniquely defined. The data consists of n observations $(y_i, \mathbf{x}_i)$ which are generated independently from distribution $P_{\mathbf{x},y}$ such that conditional distribution $P_{y|\mathbf{x}}$ is given by Eq. (1) and distribution of attribute vector $\mathbf{x}$ is $(p+1)$-dimensional with first coordinate equal to 1. We consider the case when $\mathbf{x}$ is random since in this situation behaviour of $\boldsymbol{\beta}^*$ of maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ for incorrect model specification can be more easily described (cf. definition (6) below, see however [13] for analogous development for deterministic predictors).

As a first remark note that as distribution $P_{\mathbf{x},y}$ which satisfies (1) with parameters q and $\boldsymbol{\beta}$ satisfies also (1) for parameters $\tilde{q}$ and $c\boldsymbol{\beta} + \alpha$ where $c > 0$ and $\tilde{q}(s) = q((s - \alpha)/c)$. It follows that when q is unknown only *the direction* of the vector $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_p)'$ may be possibly recovered.

Let $\mathbf{X}$ be $n \times (p+1)$ design matrix with rows $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $\mathbf{Y} = (y_1, \ldots, y_n)'$ be a response vector. Under the logistic regression model, the conditional log-likelihood function for the parameter $\boldsymbol{\gamma} \in R^{p+1}$ is

$$\begin{aligned} l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X}) &= \sum_{i=1}^{n} \{y_i \log[p(\mathbf{x}_i'\boldsymbol{\gamma})] + (1-y_i)\log[1 - p(\mathbf{x}_i'\boldsymbol{\gamma})]\} \\ &= \sum_{i=1}^{n} \{y_i \mathbf{x}_i'\boldsymbol{\gamma} - \log[1 + \exp(\mathbf{x}_i'\boldsymbol{\gamma})]\}. \end{aligned}$$

Note that we can alternatively view $l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X})$ defined above as an empirical risk corresponding to the logistic loss. Define also the score function for the parameter $\boldsymbol{\gamma} \in R^{p+1}$

$$s_n(\boldsymbol{\gamma}) = \frac{\partial l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X})}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} [y_i - p(\mathbf{x}_i'\boldsymbol{\gamma})]\mathbf{x}_i = \mathbf{X}'(\mathbf{Y} - \mathbf{p}(\boldsymbol{\gamma})), \qquad (3)$$

where $\mathbf{p}(\boldsymbol{\gamma}) = (p(\mathbf{x}_1'\boldsymbol{\gamma}), \ldots, p(\mathbf{x}_n'\boldsymbol{\gamma}))'$. The negative Hessian matrix will be denoted by

$$J_n(\boldsymbol{\gamma}) = -\frac{\partial l^2(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = \sum_{i=1}^{n} \{p(\mathbf{x}_i'\boldsymbol{\gamma})[1 - p(\mathbf{x}_i'\boldsymbol{\gamma})]\}\mathbf{x}_i\mathbf{x}_i' = \mathbf{X}'\Pi(\boldsymbol{\gamma})\mathbf{X}, \tag{4}$$

where $\Pi(\boldsymbol{\gamma}) = \text{diag}\{p(\mathbf{x}_1'\boldsymbol{\gamma})(1-p(\mathbf{x}_1'\boldsymbol{\gamma})), \ldots, p(\mathbf{x}_n'\boldsymbol{\gamma})(1-p(\mathbf{x}_n'\boldsymbol{\gamma}))\}$. Under assumption $\mathbf{E}(x_k^2) < \infty$, for $k = 1, \ldots, p$ it follows from the Law of Large Numbers that

$$n^{-1} J_n(\boldsymbol{\gamma}) \xrightarrow{P} \mathbf{E}_{\mathbf{x}}\{\mathbf{x}\mathbf{x}' p(\mathbf{x}'\boldsymbol{\gamma})[1 - p(\mathbf{x}'\boldsymbol{\gamma})]\} =: J(\boldsymbol{\gamma}). \tag{5}$$

Observe that in the case of incorrect model specification $\text{cov}[s_n(\boldsymbol{\gamma})|\mathbf{x}_1, \ldots, \mathbf{x}_n] = \sum_{i=1}^{n}\{q(\mathbf{x}_i'\boldsymbol{\gamma})[1 - q(\mathbf{x}_i'\boldsymbol{\gamma})]\}\mathbf{x}_i\mathbf{x}_i'$ is not equal to negative Hessian $J_n(\boldsymbol{\gamma})$ as in the case of correct model specification when $p(\cdot) = q(\cdot)$.

The maximum likelihood estimator (ML) $\hat{\boldsymbol{\beta}}$ of parameter $\boldsymbol{\beta}$ is defined to be

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\gamma} \in R^{p+1}} l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X}).$$

Moreover define

$$\boldsymbol{\beta}^* = \arg \min_{\boldsymbol{\gamma} \in R^{p+1}} E\{\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'\boldsymbol{\gamma})]\},$$

where

$$\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'\boldsymbol{\gamma})] = q(\mathbf{x}'\boldsymbol{\beta}) \log \frac{q(\mathbf{x}'\boldsymbol{\beta})}{p(\mathbf{x}'\boldsymbol{\gamma})} + [1 - q(\mathbf{x}'\boldsymbol{\beta})] \log \frac{1 - q(\mathbf{x}'\boldsymbol{\beta})}{1 - p(\mathbf{x}'\boldsymbol{\gamma})}$$

is the Kulback-Leibler distance from the true Bernoulli distribution with the parameter $q(\mathbf{x}'\boldsymbol{\beta})$ to the postulated one with the parameter $p(\mathbf{x}'\boldsymbol{\gamma})$. Thus $\boldsymbol{\beta}^*$ is the parameter corresponding to the logistic model closest to binary model with respect to Kullback-Leibler divergence. It follows from [16] that

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^* \tag{6}$$

Using the fact that $\partial p(\mathbf{x}'\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} = p(\mathbf{x}'\boldsymbol{\gamma})[1 - p(\mathbf{x}'\boldsymbol{\gamma})]\mathbf{x}$ it is easy to see that

$$\mathbf{E}\left[\frac{\partial \Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'\boldsymbol{\gamma})]}{\partial \boldsymbol{\gamma}}\right] = \mathbf{E}[-q(\mathbf{x}'\boldsymbol{\beta})\mathbf{x} + p(\mathbf{x}'\boldsymbol{\gamma})\mathbf{x}]$$

and

$$\mathbf{E}\left[\frac{\partial^2 \Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'\boldsymbol{\gamma})]}{\partial \boldsymbol{\gamma}\boldsymbol{\gamma}'}\right] = \mathbf{E}\{p(\mathbf{x}'\boldsymbol{\gamma})[1 - p(\mathbf{x}'\boldsymbol{\gamma})]\mathbf{x}\mathbf{x}'\}$$

is positive-semidefinite. Thus from the first of the above equations we have

$$\mathbf{E}[q(\mathbf{x}'\boldsymbol{\beta})\mathbf{x}] = \mathbf{E}[p(\mathbf{x}'\boldsymbol{\beta}^*)\mathbf{x}] = \mathbf{E}(y\mathbf{x}). \tag{7}$$

Note that as the first coordinate of $\mathbf{x}$ is equal one which corresponds to intercept, the pertaining equation is

$$\mathbf{E}[q(\mathbf{x}'\boldsymbol{\beta})] = \mathbf{E}[p(\mathbf{x}'\boldsymbol{\beta}^*)] = \mathbf{E}(y). \tag{8}$$

Using (3) and (7) we obtain

$$\begin{aligned}\operatorname{cov}\{\mathbf{E}[s_n(\boldsymbol{\beta}^*)|\mathbf{x}_1,\ldots\mathbf{x}_n]\} &= n\mathbf{E}\{\mathbf{x}\mathbf{x}'[q(\mathbf{x}'\boldsymbol{\beta}) - p(\mathbf{x}'\boldsymbol{\beta}^*)]^2\}\\ &\quad - n\mathbf{E}\{\mathbf{x}[q(\mathbf{x}'\boldsymbol{\beta}) - p(\mathbf{x}'\boldsymbol{\beta}^*)]\}\{E\{\mathbf{x}[q(\mathbf{x}'\boldsymbol{\beta}) - p(\mathbf{x}'\boldsymbol{\beta}^*)]\}\}'\\ &= n\mathbf{E}\{\mathbf{x}\mathbf{x}'[q(\mathbf{x}'\boldsymbol{\beta}) - p(\mathbf{x}'\boldsymbol{\beta}^*)]^2\}.\end{aligned}$$

We also have

$$E\{\operatorname{cov}[s_n(\boldsymbol{\beta}^*)|\mathbf{x}_1,\ldots,\mathbf{x}_n]\} = n\mathbf{E}\{\mathbf{x}\mathbf{x}'q(\mathbf{x}'\boldsymbol{\beta})[1 - q(\mathbf{x}'\boldsymbol{\beta})]\}.$$

Let $K_n(\boldsymbol{\gamma}) = \operatorname{cov}[s_n(\boldsymbol{\gamma})]$ be covariance matrix of score function $s_n(\boldsymbol{\gamma})$. From above facts we have

$$\begin{aligned}&n^{-1}K_n(\boldsymbol{\beta}^*)\\ &\quad = \mathbf{E}\left\{\mathbf{x}\mathbf{x}'\{q(\mathbf{x}'\boldsymbol{\beta})[1 - q(\mathbf{x}'\boldsymbol{\beta})] + [q(\mathbf{x}'\boldsymbol{\beta}) - p(\mathbf{x}'\boldsymbol{\beta}^*)]^2\}\right\} =: K(\boldsymbol{\beta}^*).\end{aligned} \tag{9}$$

The form of $K_n(\boldsymbol{\beta}^*)$ will be used in the proof of Lemma 2. From (6) it is also easy to see that

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\gamma}\in R^{p+1}} E\{-l(\boldsymbol{\gamma},\mathbf{Y}|\mathbf{X})\}.$$

It follows from [19] that $\boldsymbol{\beta}^*$ exists provided $0 < q(\boldsymbol{\beta}'x) < 1$ almost everywhere with respect to $P_{\mathbf{x}}$ and is unique provided $E||\mathbf{x}|| < \infty$. In the following we will always assume that $\boldsymbol{\beta}^*$ exists and is unique. In the case of correct specification, when $p(\cdot) = q(\cdot)$ we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}$. In general $\boldsymbol{\beta}^*$ may be different from $\boldsymbol{\beta}$. The most immediate example is when $q(s) = p(-s)$ which corresponds to logistic model with switched classes. In this case $\boldsymbol{\beta}^* = -\boldsymbol{\beta}$. Li and Duan [19], p. 1019 give an example when supports of β and β^* are disjoint for a loss different than logistic. Let $t^* = \{0\} \cup \{1 \le k \le p : \beta_k^* \ne 0\}$. In Sect. 4 we discuss the relationships between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ as well as between t and t^* in more detail. In Sect. 3 we give conditions under which set t^* is identified consistently. Under certain assumptions we can also have $t^* = t$ and thus identification of set t is possible.

Let us discuss the notation used in this paper. Let $m \subseteq f := \{0, 1, \ldots, p\}$ be any subset of variable indices and $|m|$ be its cardinality. Each subset m is associated with a model with explanatory variables corresponding to this subset. In the following f stands for the full model containing all available variables and by $null$ we denote model containing only intercept (indexed by 0). We denote by $\hat{\boldsymbol{\beta}}_m$ a maximum likelihood estimator calculated for model m and by $\boldsymbol{\beta}^*_m$ the minimizer of averaged Kullback-Leibler divergence when only predictors belonging to m are considered. Thus $\boldsymbol{\beta}^* = \boldsymbol{\beta}^*_f$. Moreover, $\boldsymbol{\beta}^*(m)$ stands for $\boldsymbol{\beta}^*$ restricted to m. Depending on the context these vectors will be considered as $|m|$-dimensional or as their $(p+1)$-dimensional versions augmented by zeros. We need the following fact stating that when $m \supseteq t^*$ then $\boldsymbol{\beta}^*_m$ is obtained by restricting $\boldsymbol{\beta}^*$ to m.

Lemma 1 *Let $m \supseteq t^*$ and assume $\boldsymbol{\beta}^*$ is unique. Then $\boldsymbol{\beta}^*_m = \boldsymbol{\beta}^*(m)$.*

Proof The following inequalities hold

$$\begin{aligned} E\{\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'_m\boldsymbol{\beta}^*_m)]\} &\geq E\{\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'\boldsymbol{\beta}^*)]\} \\ &= E\{\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'_m\boldsymbol{\beta}^*(m))]\}. \end{aligned}$$

From the definition of projection the above inequality is actually equality and from the uniqueness the assertion follows.

3 Consistency of Two-Step Greedy GIC Procedure

We consider the following model selection criterion

$$GIC(m) = -2l(\hat{\boldsymbol{\beta}}_m, \mathbf{Y}|\mathbf{X}) + a_n|m|,$$

where m is a given submodel containing $|m|$ variables, $\hat{\boldsymbol{\beta}}_m$ is a maximum likelihood estimator calculated for model m (augmented by zeros to p-dimensional vector) and a_n is penalty. Observe that $a_n = \log(n)$ corresponds to Bayesian Information Criterion and $a_n = 2$ corresponds to Akaike Information Criterion. GIC was considered e.g. by [22]. We would like to select a model which minimizes GIC over a family

$$\mathcal{M} := \{\{0\} \cup s : s \subseteq \{1, \ldots, p\}\},$$

i.e. the family of all submodels of f containing intercept. Denote the corresponding selector by $\hat{t}^*$. As $\mathcal{M}$ consists of 2^p models and determination of $\hat{t}^*$ requires calculation of GIC for all of them this becomes computationally unfeasible for large p. In order to restrict the space of models over which the optimal value of criterion function is sought we propose the following two-stage procedure.

Step 1. The covariates $\{1, \ldots, p\}$ are ordered with respect to the residual deviances

$$D_{f\setminus\{i_1\}f} \geq D_{f\setminus\{i_2\}f} \geq \cdots \geq D_{f\setminus\{i_p\}f}.$$

Step 2. The considered model selection criterion GIC is minimized over a family

$$\mathcal{M}_{\text{nested}} := \{\{0\}, \{0\} \cup \{i_1\}, \{0\} \cup \{i_1, i_2\}, \ldots, \{0\} \cup \{i_1, i_2, \ldots, i_p\}\}.$$

We define $\hat{t}^*_{gr}$ as the minimizer of GIC over $\mathcal{M}_{\text{nested}}$. The intuition behind the first step of the procedure is that by omitting the true regressors from the model their corresponding residual deviances are increased significantly more than when spurious ones are omitted. Thus the first step may be considered as screening of the family $\mathcal{M}$ and reducing it to $\mathcal{M}_{\text{nested}}$ by whittling away elements likely to be redundant.

The following assumption will be imposed on $P_{\mathbf{x}}$ and penalization constants a_n

(A1) $J(\boldsymbol{\beta}^*)$ is positive definite matrix.
(A2) $E(x_k^2) < \infty$, for $k = 1, \ldots, p$.
(A3) $a_n \to \infty$ and a_n/n is nonincreasing and tends to 0 as $n \to \infty$.

The main result of this section is the consistency of the greedy procedure defined above.

Theorem 1 *Under assumptions (A1)–(A3) greedy selector $\hat{t}^*_{gr}$ is consistent i.e. $P(\hat{t}^*_{gr} = t^*) \to 1$ when $n \to \infty$.*

The following two results which are of independent interest constitute the proof of Theorem 1. The first result asserts consistency of $\hat{t}^*$. This is conclusion of Proposition 4.2 (i) and (iii) in [22]. However, as the framework in the last paper is very general, it is possible to prove the assertions there under much milder assumptions without assuming e.g. that loglikelihood satisfies weak law of large numbers uniformly in β and similar assumption on J_n. Theorem 3 states that after performing the first step of the procedure relevant regressors will precede the spurious ones with probability tending to 1. Consistency of GIC in the almost sure sense was proved by [20] for deterministic regressors under some extra conditions.

Theorem 2 *Assume (A1)–(A3). Then $\hat{t}^*$ is consistent i.e.*

$$P(\hat{t}^* = t^*) = P[\min_{m \in \mathcal{M}, m \neq t^*} GIC(m) > GIC(t^*)] \to 1.$$

Consider two models j and k and denote by

$$D^n_{jk} = 2[l(\hat{\boldsymbol{\beta}}_k, \mathbf{Y}|\mathbf{X}) - l(\hat{\boldsymbol{\beta}}_j, \mathbf{Y}|\mathbf{X})] \tag{10}$$

deviance of the model k from the model j.

Theorem 3 *Assume conditions (A1)–(A2). Then for all $i \in t^* \setminus \{0\}$ and $j \notin t^* \setminus \{0\}$ we have*

$$P[D_{f\setminus\{i\}f} > D_{f\setminus\{j\}f}] \to 1, \ \textit{as } n \to \infty.$$

Proof (Theorem 1) As the number of predictors is finite and does not depend on n the assertion in Theorem 3 implies that with probability tending to one model t^* will be included in $\mathcal{M}_{\text{nested}}$. This in view of Theorem 2 yields the proof of Theorem 1.

The following lemmas will be used to prove Theorem 2. Define sequence

$$d_n^2 = \min\{[\max_{1\le i\le n} ||\mathbf{x}_i||^2]^{-1}, [\min_{k\in t^*, 1\le k\le p} (1/2)\beta_k^*]^2\}. \tag{11}$$

Remark 1 It follows from Lemma 6 that under assumptions (A2) and (A3) if $t^* \setminus 0 \neq \emptyset$ we have $nd_n^2/a_n \xrightarrow{P} \infty$.

Two lemmas below are pivotal in proving Theorem 2. The proofs are in the appendix.

Lemma 2 *Let $c \supseteq m \supseteq t^*$. Assume (A1)–(A2). Then $D_{mc} = O_P(1)$.*

Lemma 3 *Let $w \not\supseteq t^*$ and $c \supseteq t^*$. Assume (A1)–(A2). Then $P(D_{wc} > \alpha_1 nd_n^2) \to 1$ as $n \to \infty$, for some $\alpha_1 > 0$.*

Proof (Theorem 3) It follows from Lemma 3 that for $i \in t$ we have $P[D^n_{f\setminus\{i\}f} > \alpha_1 nd_n^2] \to 1$, for $\alpha_1 > 0$ and by Remark 1 $nd_n^2 \xrightarrow{P} \infty$. By Lemma 2 we have that $D_{f\setminus\{j\}f} = O_P(1)$ for $j \in t^*$, which end the proof.

Proof (Theorem 2) Consider first the case $t^* = \{0\} \cup m$, $m \neq \emptyset$. We have to show that for all models $m \in \mathcal{M}$ such that $m \neq t^*$

$$P[-2l(\hat{\boldsymbol{\beta}}_{t^*}, \mathbf{Y}|\mathbf{X}) + |t^*|a_n < -2l(\hat{\boldsymbol{\beta}}_m, \mathbf{Y}|\mathbf{X}) + |m|a_n] \to 1,$$

as $n \to \infty$ which is equivalent to $P[D_{mt^*} > a_n(|t^*| - |m|)] \to 1$. In the case of $m \not\supseteq t^*$ this follows directly from Lemma 3 and $nd_n^2/a_n \xrightarrow{P} \infty$. Consider the case of $m \supset t^*$. By Lemma 2 $D_{mt^*} = O_P(1)$. This ends the first part of the proof in view of $a_n(|t^*| - |m|) \to -\infty$. For $t^* = \{0\}$ we only consider the case $m \supset t^*$ and the assertion $P[D_{mt^*} > a_n(1 - |m|)] \to 1$ follows again from Lemma 2.

4 Interplay Between *t* and *t**

In view of the results of the previous section t^* can be consistently selected by two-step GIC procedure. As we want to choose t not t^*, the problem what is the connection between these two sets naturally arises. First we study the problem whether it is possible that t^* is $\{0\}$ whereas t does contain genuine regressors. Fortunately, the answer under some mild conditions on the distribution $P_{\mathbf{x},y}$, including monotonicity of response function q, is negative. We proceed by reexpressing the fact that $t^* = \{0\}$ in terms of conditional expectations and then showing that the obtained condition for monotone q can be satisfied only in the case when y and $\mathbf{x}$ are independent.

Let $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_p)$, $\tilde{\boldsymbol{\beta}}^* = (\beta_1^*, \ldots, \beta_p^*)$ and $\tilde{\mathbf{x}} = (x_1, \ldots, x_p)$. The first proposition (proved in the appendix) gives the simple equivalent condition for $t^* = \{0\}$.

Proposition 1 $E(\mathbf{x}|y=1) = E(\mathbf{x}|y=0)$ *if and only* $t^* = \{0\}$.

Let $f(\tilde{\mathbf{x}}|y=1)$ and $f(\tilde{\mathbf{x}}|y=0)$ be the density functions of $\tilde{\mathbf{x}}$ in classes $y=1$ and $y=0$, respectively and denote by $F(\tilde{\mathbf{x}}|y=1)$ and $F(\tilde{\mathbf{x}}|y=0)$ the corresponding probability distribution functions. Note that the above proposition in particular implies that in the logistic model for which expectations of $\mathbf{x}$ in both classes are equal we necessarily have $\tilde{\boldsymbol{\beta}} = 0$. The second proposition asserts that this is true for a general binary model under mild conditions. Thus in view of the last proposition under these conditions $t^* = \{0\}$ is equivalent to $t = \{0\}$.

Proposition 2 *Assume that q is monotone and densities $f(\tilde{\mathbf{x}}|y=1)$, $f(\tilde{\mathbf{x}}|y=0)$ exist. Then $E(\tilde{\mathbf{x}}|y=1) = E(\tilde{\mathbf{x}}|y=0)$ implies $f(\tilde{\mathbf{x}}|y=1) = f(\tilde{\mathbf{x}}|y=0)$ a.e., i.e. y and $\tilde{\mathbf{x}}$ are independent.*

Proof Define $h(\tilde{\mathbf{x}})$ as the density ratio of $f(\tilde{\mathbf{x}}|y=1)$ and $f(\tilde{\mathbf{x}}|y=0)$. Observe that as

$$h(\tilde{\mathbf{x}}) = \frac{f(\tilde{\mathbf{x}}|y=1)}{f(\tilde{\mathbf{x}}|y=0)} = \frac{P(y=0)}{P(y=1)} \frac{q(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}})}{1 - q(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}})} \tag{12}$$

we have that $h(\tilde{\mathbf{x}}) = w(\tilde{\mathbf{x}}'\tilde{\beta})$ and w is monotone.

Consider first the case $p=1$. It follows from the monotone likelihood ratio property (see [18], Lemma 2, Sect. 3) that since $h(\tilde{\mathbf{x}})$ is monotone then conditional distributions $F(\tilde{\mathbf{x}}|y=1)$ and $F(\tilde{\mathbf{x}}|y=0)$ are ordered and as their expectations are equal this implies $F(\tilde{\mathbf{x}}|y=1) = F(\tilde{\mathbf{x}}|y=0)$ and thus the conclusion for $p=1$.

For $p > 1$ assume without loss of generality that $\beta_1 \neq 0$ and consider the transformation $\mathbf{z} = (z_1, \ldots, z_p) = (\tilde{\boldsymbol{\beta}}'\tilde{\mathbf{x}}, x_2, \ldots, x_p)'$. Denote by $\tilde{f}(\mathbf{z}|y=1)$ and $\tilde{f}(\mathbf{z}|y=0)$ densities of $\mathbf{z}$ in both classes. It is easy to see that we have

$$\tilde{f}(\mathbf{z}|y=1) = \beta_1^{-1} f\left((z_1 - \beta_2 z_2 - \cdots - \beta_p z_p)/\beta_1, z_2, \ldots, z_p \middle| \, y=1\right),$$

$$\tilde{f}(\mathbf{z}|y=0) = \beta_1^{-1} f\left((z_1 - \beta_2 z_2 - \cdots - \beta_p z_p)/\beta_1, z_2, \ldots, z_p \middle| \, y=0\right)$$

and

$$\frac{\tilde{f}(\mathbf{z}|y=1)}{\tilde{f}(\mathbf{z}|y=0)} = w\left(\tilde{\boldsymbol{\beta}}'((z_1 - \beta_2 z_2, \ldots, \beta_p z_p)/\beta_1, z_2, \ldots, z_p)\right) = w(z_1). \tag{13}$$

It follows from (13) that marginal densities $\tilde{f}_1(z_1|y=1)$, $\tilde{f}_1(z_1|y=0)$ satisfy $\tilde{f}_1(z_1|y=1)/\tilde{f}_1(z_1|y=0) = w(z_1)$ and the first part of the proof yields $\tilde{f}_1(z_1|y=1) = \tilde{f}_1(z_1|y=0)$.

Thus we have for fixed z_1

$$\frac{\tilde{f}(\mathbf{z}|y=1)}{\tilde{f}(\mathbf{z}|y=0)} = \frac{\tilde{f}(z_2,\ldots,z_p|z_1,y=1)\tilde{f}_1(z_1|y=1)}{\tilde{f}(z_2,\ldots,z_p|z_1,y=0)\tilde{f}_1(z_1|y=0)} = \frac{\tilde{f}(z_2,\ldots,z_p|z_1,y=1)}{\tilde{f}(z_2,\ldots,z_p|z_1,y=0)} = w(z_1),$$

which implies that for any z_1 we have $\tilde{f}(z_2,\ldots,z_p|z_1,\mathrm{y}=1) = \tilde{f}(z_2,\ldots,z_p|z_1, y=0)$ and thus $\tilde{f}(\mathbf{z}|y=1) = \tilde{f}(\mathbf{z}|y=0)$ and consequently $f(\tilde{\mathbf{x}}|y=1) = f(\tilde{\mathbf{x}}|y=0)$ which ends the proof.

Observe now that in view of (12) if $f(\tilde{\mathbf{x}}|y=1) = f(\tilde{\mathbf{x}}|y=0)$ then $q(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}})$ is constant and thus $\tilde{\boldsymbol{\beta}} = 0$ if $1, x_1, \ldots, x_p$ are linearly independent with probability 1 i.e. $\mathbf{x}'\mathbf{b} = b_0$ a.e. implies that $\mathbf{b} = 0$ (or equivalently that $\Sigma_{\mathbf{x}} > 0$). Thus we obtain

Theorem 4 *If q is monotone and not constant and $1, x_1, \ldots, x_p$ are linearly independent with probability 1 then $t^* = \{0\}$ is equivalent to $t = \{0\}$ or, $\tilde{\boldsymbol{\beta}}^* \neq 0$ is equivalent to $\tilde{\boldsymbol{\beta}} \neq \mathbf{0}$.*

Now we address the question when $t = t^*$. The following theorem has been proved in [21], see also [19] for a simple proof based on generalized Jensen inequality.

Theorem 5 *Assume that $\boldsymbol{\beta}^*$ is uniquely defined and there exist $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1 \in R^p$ such that*

(R) $E(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}'\boldsymbol{\beta} = z) = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 z.$

Then $\tilde{\boldsymbol{\beta}}^* = \eta\tilde{\boldsymbol{\beta}}$, for some $\eta \in R$.

It is well known that Rudd's condition (R) is satisfied for eliptically contoured distributions. In particular multivariate normal distribution satisfies this property (see e.g. [19], Remark 2.2). The case when $\eta \neq 0$ plays an important role as it follows from the assertion of Theorem 5 that then $t^* = t$. Note that in many statistical problems we want to consistently estimate the direction of vector $\boldsymbol{\beta}$ and not its length. This is true for many classification methods when we look for direction such that projection on this direction will give maximal separation of classes. Theorem 4 implies that under its conditions η in the assertion of Theorem 5 is not equal zero. Thus we can state

Corollary 1 *Assume (A1)–(A3), (R) and conditions of Theorem 4. Then*

$$P(\hat{t}^*_{gr} = t) \to 1$$

i.e. two-stage greedy GIC is consistent for t.

Proof Under (R) it follows from Theorem 5 that $\tilde{\boldsymbol{\beta}}^* = \eta\tilde{\boldsymbol{\beta}}$ and as q is monotone and not constant it follows from Theorem 4 that $\eta \neq 0$ and thus $t = t^*$. This implies the assertion in view of Theorem 2.

In the next section by means of numerical experiments we will indicate that magnitude of η plays an important role for probability of correct selection. In particular we will present examples showing that when regressors are jointly normal and thus Ruud's condition is satisfied, probability of correct selection of t by two-step greedy GIC can be significantly larger under misspecification than under correct specification.

The analogous result to Corollary 1 follows for $\hat{t}^*$ when GIC is minimized over the whole family of 2^p models.

The important consequence of Theorem 4 is that power of significance test will increase to 1 when there is dependence of y on $\mathbf{x}$ even when logistic model is misspecified and critical region is constructed for such model. Namely, consider significance test for $H_0 : \tilde{\boldsymbol{\beta}} = 0$ with critical region

$$\mathcal{C}_{1-\alpha} = \{D_{null,\hat{t}^*_{gr}} > \chi^2_{|\hat{t}^*_{gr}|-1,1-\alpha}\} \tag{14}$$

where $\chi^2_{k,1-\alpha}$ is quantile of order $1-\alpha$ of chi-squared distribution with k degrees of freedom. Observe that if $p = q$ it follows from Theorem 2 and [12] that under null hypothesis $P(C_{1-\alpha}|H_0) \to \alpha$ what explains the exact form of the threshold of the rejection region when the logistic model is fitted. We have

Corollary 2 *Assume that conditions of Theorem 4 are satisfied and $\tilde{\boldsymbol{\beta}} \neq 0$. Consider test of $H_0 : \tilde{\boldsymbol{\beta}} = \mathbf{0}$ against $H_1 : \tilde{\boldsymbol{\beta}} \neq \mathbf{0}$ with critical region $\mathcal{C}_{1-\alpha}$ defined in (14). Then the test is consistent i.e. $P(D_{null,\hat{t}^*_{gr}} \in C_{1-\alpha}|H_1) \to 1$.*

Observe that if $\tilde{\boldsymbol{\beta}}^* \neq \mathbf{0}$ then in view of Remark 1 $nd_n^2 \to \infty$. Then the main results and Lemma 3 imply that when $\tilde{\boldsymbol{\beta}}^* \neq \mathbf{0}$ $P[D_{null,\hat{t}^*_{gr}} > \chi^2_{|\hat{t}^*_{gr}|-1,1-\alpha}] \to 1$ for any $\alpha > 0$ and the test is consistent. But in view of Theorem 4 $\tilde{\boldsymbol{\beta}}^* \neq \mathbf{0}$ is implied by $\tilde{\boldsymbol{\beta}} \neq 0$.

5 Numerical Experiments

In this section we study how the incorrect model specification affects the model selection and testing procedures, in particular how it influences probability of correct model selection, positive selection rate, false discovery rate and power of a test of significance. In the case when attributes are normally distributed we investigate how these measures depend on proportionality constant η appearing in Rudd's theorem.

Recall that t denotes the minimal true model. Convention that $\boldsymbol{\beta}_t$ is subvector of $\boldsymbol{\beta}$ corresponding to t is used throughout. We consider the following list of models.

(M1) $t = \{10\}$, $\beta_t = 0.2$,
(M2) $t = \{2, 4, 5\}$, $\boldsymbol{\beta}_t = (1, 1, 1)'$,
(M3) $t = \{1, 2\}$, $\boldsymbol{\beta}_t = (0.5, 0.7)'$,
(M4) $t = \{1, 2\}$, $\boldsymbol{\beta}_t = (0.3, 0.5)'$,
(M5) $t = \{1, \ldots, 8\}$, $\boldsymbol{\beta}_t = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)'$.

Models (M3)–(M5) above are considered in [9]. The number of all potential attributes is initially set to be $p = 15$ so the proportion of relevant variables varies from 6.66 % (for model M1) to 53.33 % (for model M5). Recall that $q(\cdot)$ denotes a true response function, i.e. for a given $\mathbf{x}$, y is generated from Bernoulli distribution with success probability $q(\mathbf{x}'\boldsymbol{\beta})$. The logistic model defined in (2) is fitted. Let $F_{N(0,1)}(\cdot)$ denote distribution function of standard normal random variable and $F_{Cauchy(u,v)}(\cdot)$ distribution function of Cauchy distribution with location u and scale v. In the case of incorrect model specification, the following response functions are considered:

$$q_1(s) = F_{N(0,1)}(s) \quad \text{(Probit model)},$$

$$q_2(s) = \begin{cases} F_{N(0,1)}(s) & \text{for } F_{N(0,1)}(s) \in (0.1, 0.8) \\ 0.1 & \text{for } F_{N(0,1)}(s) \leq 0.1 \\ 0.8 & \text{for } F_{N(0,1)}(s) \geq 0.8, \end{cases}$$

$$q_3(s) = \begin{cases} F_{N(0,1)}(s) & \text{for } F_{N(0,1)}(s) \in (0.2, 0.7) \\ 0.2 & \text{for } F_{N(0,1)}(s) \leq 0.2 \\ 0.7 & \text{for } F_{N(0,1)}(s) \geq 0.7, \end{cases}$$

$$q_4(s) = \begin{cases} F_{N(0,1)}(s) & \text{for } |s| > 1 \\ 0.5 + 0.5\cos[4\pi F_{N(0,1)}(s)]F_{N(0,1)}(s) & \text{for } |s| \leq 1, \end{cases}$$

$$q_5(s) = F_{Cauchy(0,1)}(s),$$

$$q_6(s) = F_{Cauchy(0,2)}(s),$$

Studied response functions are shown in Fig. 1. Dashed line there corresponds to fitted logistic response function $p(\cdot)$.

We consider two distributions of attributes, in both cases attributes are assumed to be independent. In the first scenario x_j have $N(0, 1)$ distribution and in the second x_j are generated from Gaussian mixture $0.95N(0, 1) + 0.05N(5, 1)$. Thus in the first case condition (R) of Theorem 5 is satisfied. This implies $\tilde{\boldsymbol{\beta}}^* = \eta\tilde{\boldsymbol{\beta}}$, for some $\eta \in R$. One of our main goals is to investigate how the value of η affects the performance of model selection and testing procedures.

Recall that although Rudd's condition is a property of distribution of predictors and $\boldsymbol{\beta}$ it follows from definition of $\boldsymbol{\beta}^*$ that η depends on the model as well as on misspecified response $q(\cdot)$. Table 1 shows values of estimated proportionality constant η, denoted by $\hat{\eta}$. To calculate $\hat{\eta}$, for each variable $k \in t$, the value $\hat{\beta}_k/\beta_k$, where $\hat{\boldsymbol{\beta}}$ is based on $n = 10^6$ observations is computed and then the values are averaged over all attributes. The first column corresponds to $\eta = 1$ and it allows to gauge the variability of $\hat{\eta}$. Note also that the smallest value of $\hat{\eta}$ equal 0.52 and the second largest (equal 1.74) are obtained for the model M2 and responses q_6 and q_1, respectively. It follows that in the first case estimated $\boldsymbol{\beta}$ is on average two times smaller than the true one and around 1.7 times larger in the second case. Observe also that when $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ are approximately proportional, for $q(s)$ such that $q(s) > p(s)$ for $s > 0$ we can expect that $\hat{\boldsymbol{\beta}} > \boldsymbol{\beta}$ as we try to match $q(\mathbf{x}_i'\boldsymbol{\beta})$ with $p(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$.

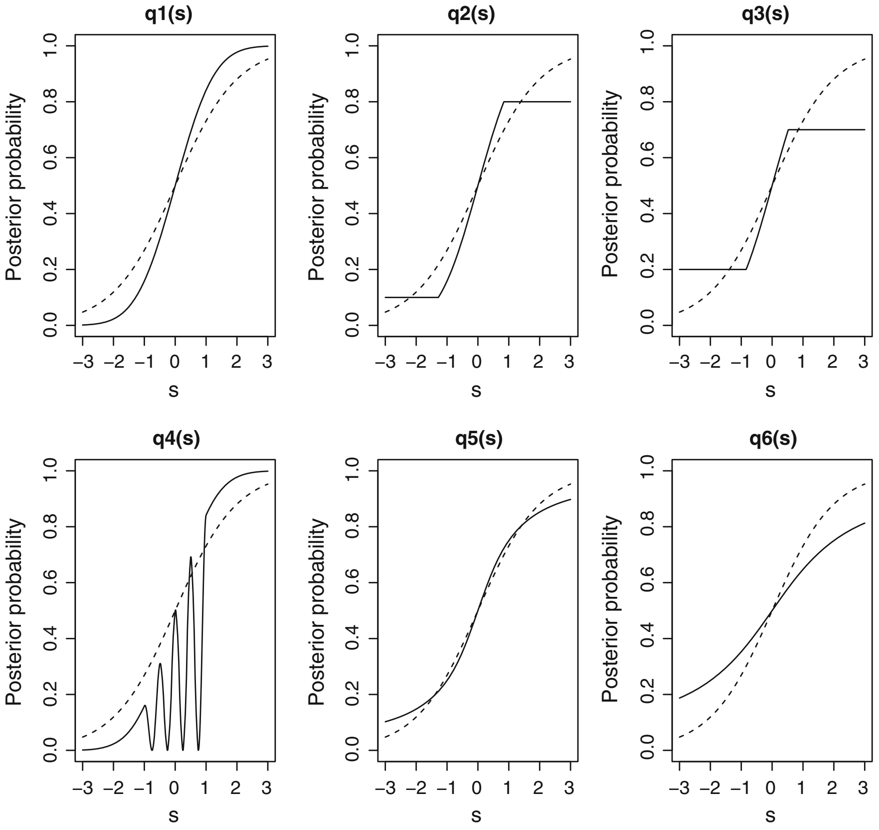


Fig. 1 Responses functions. *Dashed line* corresponds to fitted logit model $p(\cdot)$

Table 1 Values of $\hat{\eta}$ for considered models

Model	$p(\cdot)$	$q_1(\cdot)$	$q_2(\cdot)$	$q_3(\cdot)$	$q_4(\cdot)$	$q_5(\cdot)$	$q_6(\cdot)$
M1	0.988	1.642	1.591	1.591	0.788	1.241	0.651
M2	1.005	1.741	0.863	0.537	1.735	0.874	0.522
M3	0.993	1.681	1.352	0.968	1.524	1.045	0.580
M4	1.005	1.644	1.510	1.236	1.293	1.140	0.610
M5	1.013	1.779	0.897	0.552	1.724	0.879	0.532

This results in $\hat{\eta} > 1$. Thus as expected for q_1, $\hat{\eta}$ is greater than 1, whereas for q_6 it is smaller than 1.

It is noted in [2] (Sect. 4.2) that the probit function can be approximated by the scaled logit function as $q_1(s) \approx p(a \cdot s)$, where the scaling constant $a = \sqrt{8/\pi} \approx 1.6$ is chosen so that the derivatives of the two curves are equal for $s = 0$. Observe that constant a is very close to $\hat{\eta}$ calculated for q_1 (see Table 1).

In order to select the final model we use the two-step greedy procedure with Bayesian Information Criterion (BIC) described in Sect. 3. All fitted models include intercept.

Let $\hat{t}^*$ denote the model selected by a given selection criterion. As the measures of performance we use the following indices:

- probability of correct model selection (CS): $P(\hat{t}^* = t)$,
- positive selection rate (PSR): $\mathbf{E}(|\hat{t}^* \cap t|/|t|)$,
- false discovery rate (FDR): $\mathbf{E}(|\hat{t}^* \setminus t|/|\hat{t}^*|)$,
- power of significance test (POWER): $P(D_{null,\hat{t}^*} \in C_{1-\alpha}|H_1)$, where $\mathcal{C}_{1-\alpha}$ is critical region and H_1 corresponds to models M1–M5. Level $\alpha = 0.05$ was adopted throughout.

Empirical versions of the above measures are calculated and the results are averaged over 200 simulations. In the case of difficult models containing several predictors with small contributions CS can be close to zero and thus PSR and FDR are much more revealing measures of effectiveness. Observe that PSR is an average fraction of correctly chosen variables with respect to all significant ones whereas FDR measures a fraction of false positives (selected variables which are not significant) with respect to all chosen variables. Thus PSR $= 1$ means that all significant variables are included in the chosen model whereas FDR $= 0$ corresponds to the case when no spurious covariates are present in the final model. Instead of using critical region based on asymptotic distribution defined in (14) for which the significance level usually significantly exceeded assumed one, Monte Carlo critical value is calculated. For a given n and p 10000 datasets from null model are generated, for each one $\hat{t}^*$ and $D_{null,\hat{t}^*}$ is computed and this yields distribution of $D_{null,\hat{t}^*}$. The critical value is defined as empirical quantile of order $(1-\alpha)$ for $D_{null,\hat{t}^*}$.

Table 2 shows the results for $n = 200$. The highlighted values are maximal value in row (minimal values in case of FDR) and the last column pertains to maximal standard deviation in row. Observe that the type of response function influences greatly all considered measures of performance. Values of POWER are mostly larger than CS as detection of at least one significant variable usually leads to rejection of the null hypothesis. The most significant differences are observed for model M5 for which it is difficult to identify all significant variables as some coefficients are close to zero but it is much easier to reject the null model. However, when there is only one significant variable in the model, the opposite may be true as it happens for model M1. Note also that CS, PSR and POWER are usually large for large $\hat{\eta}$. To make this point more clear Fig. 2 shows the dependence of CS, PSR, POWER on $\hat{\eta}$. Model M1 is not considered for this graph as it contains only one significant predictor. In the case of CS, PSR and POWER monotone dependence is evident. However FDR is unaffected by the value of η which is understandable in view of its definition.

Table 3 shows the results for $n = 200$ when attributes x_j are generated from Gaussian mixture $0.95N(0,1) + 0.05N(5,1)$. Observe that the greatest impact of the change of $\mathbf{x}$ on CS occurs for truncated probit responses q_2 and q_3 for which in the case of M2–M5 CS drops dramatically. The change affects also PSR but to a lesser extent.

Table 2 CS, PSR, FDR and POWER for $x_j \sim N(0, 1)$ with $n = 200$, $p = 15$

Model		$p(\cdot)$	$q_1(\cdot)$	$q_2(\cdot)$	$q_3(\cdot)$	$q_4(\cdot)$	$q_5(\cdot)$	$q_6(\cdot)$	max sd
M1	CS	0.100	**0.410**	**0.410**	0.400	0.070	0.190	0.060	0.035
	PSR	0.170	**0.530**	**0.530**	0.520	0.110	0.300	0.080	0.036
	FDR	0.218	0.198	0.198	0.198	**0.142**	0.234	0.243	0.030
	POWER	0.080	**0.200**	**0.200**	**0.200**	0.110	0.120	0.040	0.028
M2	CS	0.820	0.760	0.850	0.550	0.770	**0.870**	0.590	0.035
	PSR	**1.000**	**1.000**	**1.000**	0.860	**1.000**	**1.000**	0.867	0.016
	FDR	0.050	0.072	0.040	0.051	0.061	**0.038**	0.064	0.011
	POWER	**1.000**	**1.000**	**1.000**	0.970	**1.000**	**1.000**	0.970	0.012
M3	CS	0.680	**0.790**	0.760	0.670	0.680	0.660	0.250	0.034
	PSR	0.920	**0.995**	0.975	0.910	0.985	0.940	0.590	0.023
	FDR	0.068	0.073	0.082	**0.060**	0.103	0.095	0.087	0.013
	POWER	0.980	**1.000**	**1.000**	0.950	**1.000**	0.990	0.550	0.035
M4	CS	0.300	**0.700**	0.680	0.440	0.380	0.380	0.050	0.035
	PSR	0.650	**0.940**	0.920	0.795	0.740	0.765	0.310	0.023
	FDR	0.130	0.078	**0.073**	0.113	0.140	0.103	0.153	0.021
	POWER	0.700	**1.000**	0.990	0.890	0.870	0.830	0.290	0.033
M5	CS	0.000	0.090	0.010	0.000	**0.110**	0.000	0.000	0.022
	PSR	0.647	**0.821**	0.601	0.391	0.815	0.595	0.372	0.012
	FDR	0.033	0.031	0.034	0.047	**0.024**	0.038	0.068	0.010
	POWER	**1.000**	**1.000**	**1.000**	0.950	**1.000**	**1.000**	0.930	0.018

To investigate this effect further we consider the probit function truncated at levels c and $1 - c$

$$q_7(s) = \begin{cases} F_{N(0,1)}(s) & \text{for } F_{N(0,1)}(s) \in (c, 1-c) \\ 0.2 & \text{for } F_{N(0,1)}(s) \le c \\ 0.7 & \text{for } F_{N(0,1)}(s) \ge 1-c, \end{cases}$$

which is a generalization of q_2 and q_3. Figure 7 shows how parameter c influences CS, PSR and FDR when the response is generated from q_7 and attributes are generated from Gaussian mixture $0.95N(0, 1) + 0.05N(5, 1)$.

To illustrate the result concerning the consistency of greedy two-step model selection procedure stated in Corollary 1 we made an experiment in which dependency on n is investigated. Figures 3 and 4 show considered measures of performance with respect to n for models M4 and M5. Somehow unexpectedly in some situations the results for incorrect model specification are better than for the correct specification, e.g. for model (M4) CS is larger for q_1, q_2 and q_4 than for $q(\cdot) = p(\cdot)$ (cf. Fig. 3). The results for q_6 are usually significantly worse than for p, which is related to the fact that $\hat{\eta}$ for this response is small (see again Table 1). Observe also that the type of response function clearly affects the PSRs whereas FDRs are similar in all cases.

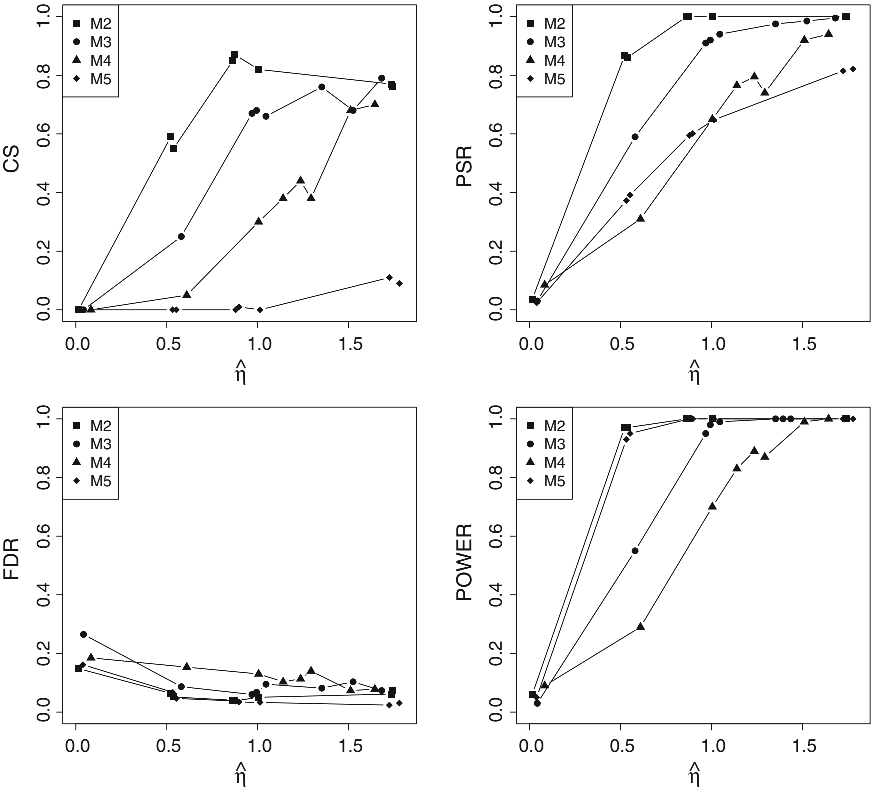


Fig. 2 CS, PSR, FDR, POWER versus $\hat{\eta}$ for $n = 200$, $p = 15$. Each point corresponds to different response function

Figure 5 shows how the power of the test of significance for the selected model and for the full model depends on the value of coefficient corresponding to the relevant variable in model M1. We see that for both correct and incorrect specification the power for selected model is slightly larger than for the full model for sufficiently large value of coefficient β_{10}. The difference is seen for smaller values of $\boldsymbol{\beta}$ in case of misspecification.

Finally we analysed how the number of potential attributes p influences the performance measures. The results shown in Fig. 6 for model M1 and $n = 500$ indicate that FDR increases significantly when spurious variables are added to the model. At the same time CS decreases when p increases, however, PSR is largely unaffected.

In conclusion we have established that when predictors are normal quality of model selection and power of the deviance test depend on the magnitude of Rudd's constant η. When $\eta > 1$ one can expect better results than for correct specification. Moreover, values of CS, PSR and POWER depend monotonically on η.

Table 3 CS, PSR, FDR and POWER for $x_j \sim 0.95N(0, 1) + 0.05N(5, 1)$ with $n = 200$, $p = 15$

Model		$p(\cdot)$	$q_1(\cdot)$	$q_2(\cdot)$	$q_3(\cdot)$	$q_4(\cdot)$	$q_5(\cdot)$	$q_6(\cdot)$	max sd
M1	CS	0.140	**0.540**	0.490	0.370	0.270	0.220	0.060	0.036
	PSR	0.220	**0.700**	0.670	0.490	0.330	0.320	0.110	0.036
	FDR	0.403	0.263	0.270	**0.233**	0.452	0.344	0.245	0.034
	POWER	0.220	**0.460**	0.450	0.240	0.340	0.260	0.090	0.035
M2	CS	**0.790**	0.730	0.180	0.050	0.780	0.720	0.350	0.034
	PSR	0.993	**1.000**	0.943	0.573	**1.000**	0.977	0.777	0.021
	FDR	**0.052**	0.070	0.278	0.227	0.056	0.084	0.094	0.016
	POWER	**1.000**	**1.000**	0.990	0.740	**1.000**	**1.000**	0.980	0.031
M3	CS	0.600	**0.740**	0.140	0.090	0.700	0.440	0.140	0.035
	PSR	0.925	**1.000**	0.915	0.725	0.990	0.855	0.600	0.021
	FDR	0.103	**0.095**	0.338	0.283	0.106	0.169	0.163	0.019
	POWER	**1.000**	**1.000**	0.990	0.840	**1.000**	**1.000**	0.790	0.029
M4	CS	0.330	**0.670**	0.120	0.040	0.410	0.210	0.010	0.035
	PSR	0.690	**0.920**	0.700	0.620	0.800	0.685	0.385	0.020
	FDR	0.148	**0.077**	0.235	0.230	0.127	0.147	0.248	0.027
	POWER	0.950	**1.000**	0.930	0.760	**1.000**	0.890	0.460	0.035
M5	CS	0.010	**0.140**	0.000	0.000	0.070	0.000	0.000	0.025
	PSR	0.641	**0.834**	0.338	0.194	0.792	0.573	0.324	0.011
	FDR	**0.013**	0.020	0.188	0.185	0.017	0.034	0.054	0.015
	POWER	**1.000**	**1.000**	0.970	0.720	**1.000**	**1.000**	0.960	0.032

In addition to tests on simulated data we performed an experiment on real data. We used Indian Liver Patient Dataset publicly available at UCI Machine Learning Repository [1]. This data set contains 10 predictors: age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. The binary response indicates whether the patient has a liver disease or not. Our aim was to use real explanatory variables describing the patients to generate an artificial response from different response functions. This can mimic the situation in which the liver disease cases follow some unknown distribution depending on explanatory variables listed above. We applied the following procedure. Predictors chosen by stepwise backward selection using BIC were considered. Estimators pertaining to 3 chosen variables (1st-age, 4th-direct Bilirubin and 6th-albumin) are treated as new true parameters corresponding to significant variables whereas the remaining variables are treated as not significant ones. Having the new parameter $\boldsymbol{\beta}$ and vectors of explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in the data we generate new $y_1, \ldots, y_n$ using considered response functions $p, q_1, \ldots, q_6$.

Table 4 shows fraction of simulations in which the given variable was selected to the final model when the two-step procedure was applied. Note that this measure is less restrictive than CS used in previous experiments. Observe that the choice of response function affects the probabilities, e.g. direct Bilirubin is chosen in 80 %

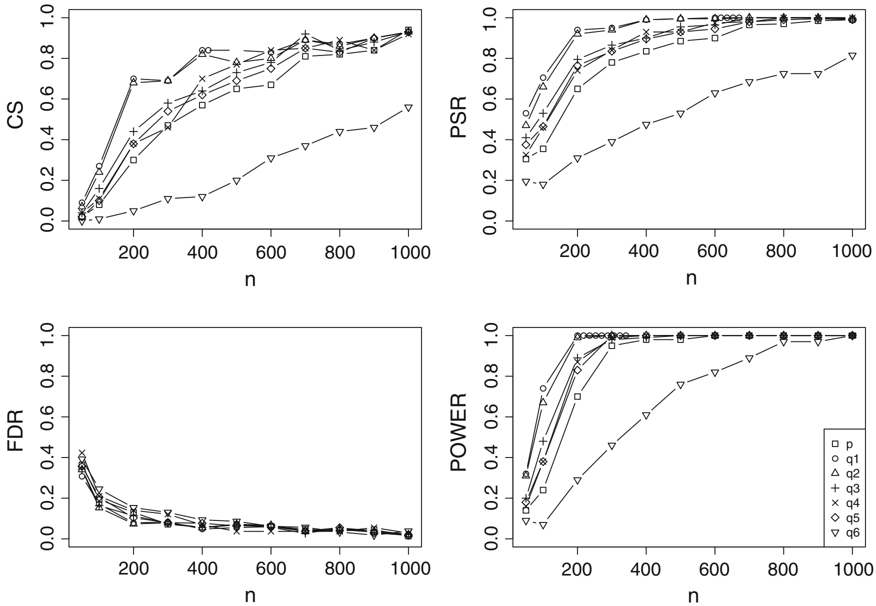


Fig. 3 CS, PSR, FDR, POWER versus *n* for model (M4), $p = 15$. Note change of the scale for FDR

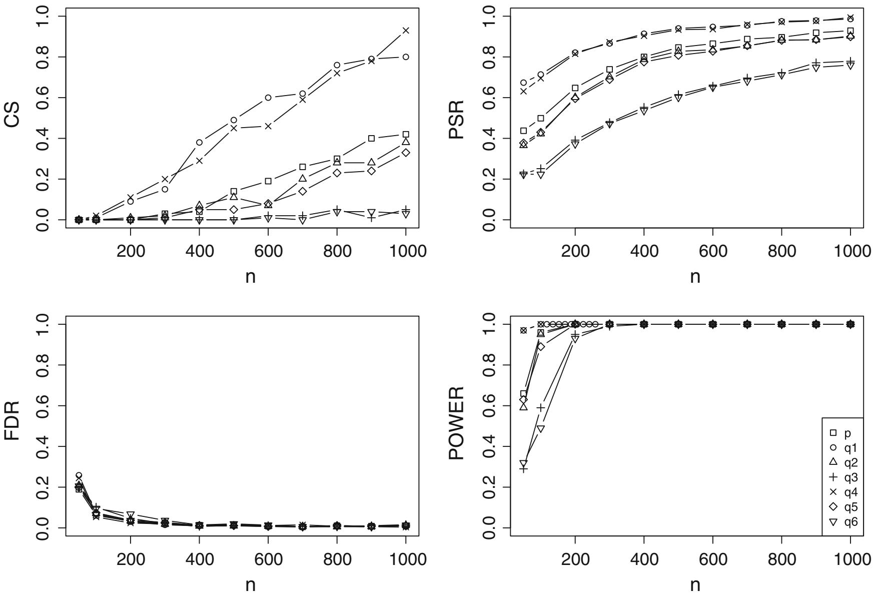


Fig. 4 CS, PSR, FDR, POWER versus *n* for model (M5), $p = 15$. Note change of the scale for FDR

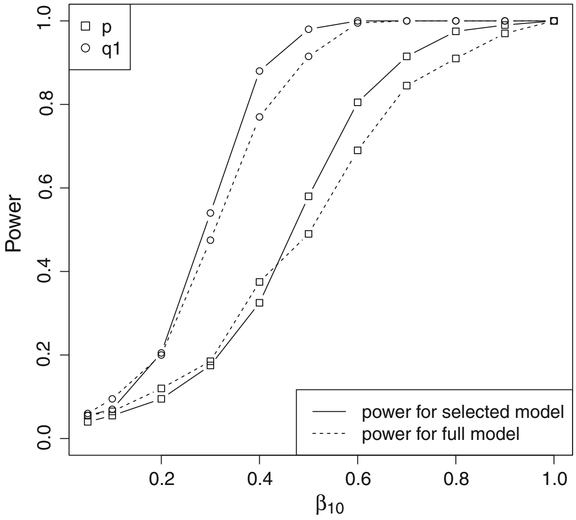


Fig. 5 Power versus β_{10} for selected model and full model, with $n = 200$, $p = 15$

simulations for correct specification and only in 12 % simulations for q_3. The significant variables are most often chosen to the final model for p and q_1. It is seen that direct Bilirubin is less likely to be selected in the case of most of the considered response functions (Fig. 7).

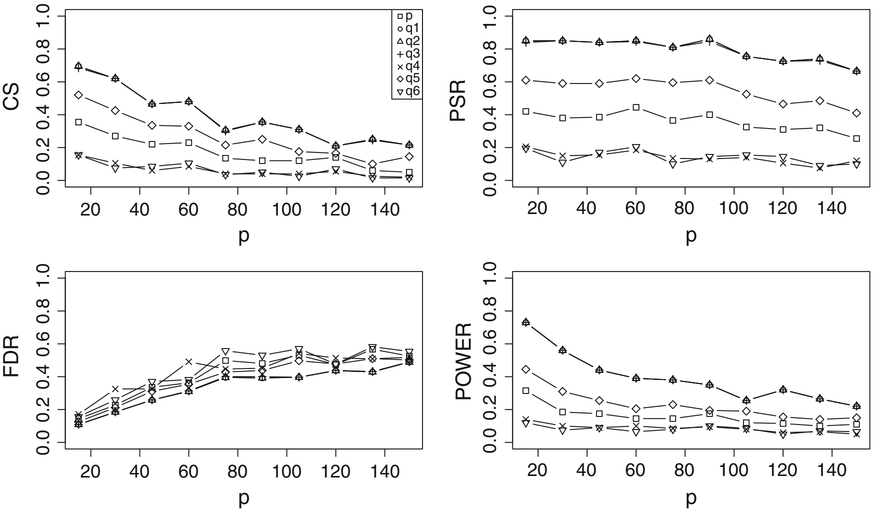


Fig. 6 CS, PSR, FDR, POWER versus p for model M1 with $n = 500$

Table 4 Probabilities of selecting variables to the final model for Indian liver patient dataset

Relevant variable	$\boldsymbol{\beta}$	p	q_1	q_2	q_3	q_4	q_5	q_6
1	−0.02	0.95	1.00	1.00	0.88	0.87	0.95	0.62
0	0.00	0.12	0.13	0.20	0.11	0.09	0.09	0.11
0	0.00	0.23	0.23	0.16	0.07	0.18	0.19	0.27
1	−0.67	0.80	0.77	0.36	0.12	0.30	0.60	0.63
0	0.00	0.11	0.15	0.26	0.08	0.17	0.10	0.17
1	−0.02	1.00	1.00	0.44	0.10	0.95	0.84	0.72
0	0.00	0.17	0.17	0.27	0.05	0.09	0.22	0.19
0	0.00	0.23	0.16	0.13	0.01	0.08	0.15	0.16
0	0.00	0.28	0.15	0.06	0.02	0.08	0.18	0.17
0	0.00	0.22	0.14	0.06	0.04	0.10	0.16	0.12

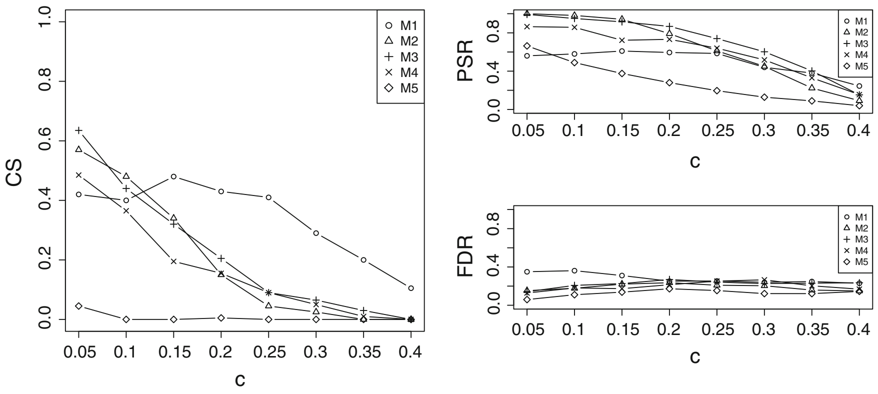


Fig. 7 CS, PSR, FDR versus c for q_7, $x_j \sim 0.95N(0, 1) + 0.05N(5, 1)$, $n = 200$ and $p = 15$

Appendix A: Auxiliary Lemmas

This section contains some auxiliary facts used in the proofs. The following theorem states the asymptotic normality of maximum likelihood estimator.

Theorem 6 *Assume (A1) and (A2). Then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{d} N(0, J^{-1}(\boldsymbol{\beta}^*)K(\boldsymbol{\beta}^*)J^{-1}(\boldsymbol{\beta}^*))$$

where J and K are defined in (5) and (9), respectively.

The above Theorem is stated in [11] (Theorem 3.1) and in [16] ((2.10) and Sect. $5B$).

Lemma 4 *Assume that* $\max_{1\le i\le n}|\mathbf{x}_i'(\boldsymbol{\gamma}-\boldsymbol{\beta})|\le C$ *for some* $C>0$ *and some* $\boldsymbol{\gamma}\in R^{p+1}$. *Then for any* $\mathbf{c}\in R^{p+1}$

$$\exp(-3C)\mathbf{c}'J_n(\boldsymbol{\beta})\mathbf{c}\le\mathbf{c}'J_n(\boldsymbol{\gamma})\mathbf{c}\le\exp(3C)\mathbf{c}'J_n(\boldsymbol{\beta})\mathbf{c},\quad a.e.$$

Proof It suffices to show that for $i=1,\dots,n$

$$\exp(-3C)p(\mathbf{x}_i'\boldsymbol{\beta})[1-p(\mathbf{x}_i'\boldsymbol{\beta})]\le p(\mathbf{x}_i'\boldsymbol{\gamma})[1-p(\mathbf{x}_i'\boldsymbol{\gamma})]\le\exp(3C)p(\mathbf{x}_i'\boldsymbol{\beta})[1-p(\mathbf{x}_i'\boldsymbol{\beta})].$$

Observe that for $\boldsymbol{\gamma}$ such that $\max_{i\le n}|\mathbf{x}_i'(\boldsymbol{\gamma}-\boldsymbol{\beta})|\le C$ there is

$$\frac{p(\mathbf{x}_i'\boldsymbol{\gamma})[1-p(\mathbf{x}_i'\boldsymbol{\gamma})]}{p(\mathbf{x}_i'\boldsymbol{\beta})[1-p(\mathbf{x}_i'\boldsymbol{\beta})]}=e^{\mathbf{x}_i'(\boldsymbol{\gamma}-\boldsymbol{\beta})}\left[\frac{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\gamma}}}\right]^2\ge e^{-C}\left[\frac{e^{-\mathbf{x}_i'\boldsymbol{\beta}}+1}{e^{-\mathbf{x}_i'\boldsymbol{\beta}}+e^{C}}\right]^2\ge e^{-3C}. \tag{15}$$

By replacing $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in (15) we obtain the upper bound for $\mathbf{c}'J_n(\boldsymbol{\gamma})\mathbf{c}$.

Lemma 5 *Assume (A1) and (A2). Then* $l(\hat{\boldsymbol{\beta}},\mathbf{Y}|\mathbf{X})-l(\boldsymbol{\beta}^*,\mathbf{Y}|\mathbf{X})=O_P(1)$.

Proof Using Taylor expansion we have for some $\bar{\boldsymbol{\beta}}$ belonging to the line segment joining $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$

$$l(\hat{\boldsymbol{\beta}},\mathbf{Y}|\mathbf{X})-l(\boldsymbol{\beta}^*,\mathbf{Y}|\mathbf{X})=\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)'[J_n(\bar{\boldsymbol{\beta}})/n]\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)/2, \tag{16}$$

Define set $A_n=\{\boldsymbol{\gamma}:||\boldsymbol{\gamma}-\boldsymbol{\beta}^*||\le s_n\}$, where s_n is an arbitrary sequence such that $ns_n^2\to 0$. Using Schwarz and Markov inequalities we have for any $C>0$

$$\begin{aligned}P[\max_{i\le i\le n}|\mathbf{x}_i'(\boldsymbol{\gamma}-\boldsymbol{\beta}^*)|>C]&\le P[\max_{1\le i\le n}||\mathbf{x}_i||s_n>C]\\&\le n\max_{i\le i\le n}P[||\mathbf{x}_i||>Cs_n^{-1}]\le C^{-2}ns_n^2\mathbf{E}(||\mathbf{x}||^2)\to 0.\end{aligned}$$

Thus using Lemma 4 the quadratic form in (16) is bounded with probability tending to 1 from above by

$$\exp(3C)\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)'[J_n(\boldsymbol{\beta}^*)/n]\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)/2,$$

which is $O_P(1)$ as $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)=O_P(1)$ in view of Theorem 6 and $n^{-1}J_n(\boldsymbol{\beta}^*)\xrightarrow{P}J(\boldsymbol{\beta}^*)$.

A.1 Proof of Lemma 2

As $\boldsymbol{\beta}_m^* = \boldsymbol{\beta}_c^*$ we have for $c \supseteq m \supseteq t^*$

$$l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}|\mathbf{X}) - l(\hat{\boldsymbol{\beta}}_m, \mathbf{Y}|\mathbf{X}) = [l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}|\mathbf{X}) - l(\boldsymbol{\beta}_c^*, \mathbf{Y}|\mathbf{X})] + [l(\boldsymbol{\beta}_m^*, \mathbf{Y}|\mathbf{X}) - l(\hat{\boldsymbol{\beta}}_m|\mathbf{X}, \mathbf{Y})],$$

which is $O_P(1)$ in view of Remark 1 and Lemma 5.

A.2 Proof of Lemma 3

The difference $l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}|\mathbf{X}) - l(\hat{\boldsymbol{\beta}}_w, \mathbf{Y}|\mathbf{X})$ can be written as

$$[l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}|\mathbf{X}) - l(\boldsymbol{\beta}^*, \mathbf{Y}|\mathbf{X})] + [l(\boldsymbol{\beta}^*, \mathbf{Y}|\mathbf{X}) - l(\hat{\boldsymbol{\beta}}_w|\mathbf{X}, \mathbf{Y})]. \tag{17}$$

It follows from Lemma 5 and Remark 1 that the first term in (17) is $O_P(1)$. We will show that the probability that the second term in (17) is greater or equal $\alpha_1 n d_n^2$, for some $\alpha_1 > 0$ tends to 1. Define set $A_n = \{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\beta}^*|| \leq d_n\}$. Using the Schwarz inequality we have

$$\sup_{\boldsymbol{\gamma} \in A_n} \max_{i \leq n} |\mathbf{x}_i'(\boldsymbol{\gamma} - \boldsymbol{\beta}^*)| < \max_{1 \leq i \leq n} ||\mathbf{x}_i|| d_n \leq 1, \tag{18}$$

with probability one. Define $H_n(\boldsymbol{\gamma}) = l(\boldsymbol{\beta}^*, \mathbf{Y}|\mathbf{X}) - l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X})$. Note that $H(\boldsymbol{\gamma})$ is convex and $H(\boldsymbol{\beta}^*) = 0$. For any incorrect model w, in view of definition (11) of d_n, we have $\hat{\boldsymbol{\beta}}_w \notin A_n$ for sufficiently large n. Thus it suffices to show that $P(\inf_{\boldsymbol{\gamma} \in \partial A_n} H_n(\boldsymbol{\gamma}) > \alpha_1 n d_n^2) \to 1$, as $n \to \infty$, for some $\alpha_1 > 0$. Using Taylor expansion for some $\bar{\boldsymbol{\gamma}}$ belonging to the line segment joining $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}^*$

$$l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X}) - l(\boldsymbol{\beta}^*, \mathbf{Y}|\mathbf{X}) = (\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' s_n(\boldsymbol{\beta}^*) - (\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' J_n(\bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma} - \boldsymbol{\beta}^*)/2$$

and the last convergence is implied by

$$P[\sup_{\boldsymbol{\gamma} \in \partial A_n} (\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' s_n(\boldsymbol{\beta}^*) > \inf_{\boldsymbol{\gamma} \in \partial A_n} (\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' J_n(\bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma} - \boldsymbol{\beta}^*)/2 - \alpha_1 n d_n^2] \to 0. \tag{19}$$

It follows from Lemma 4 and (18) that for $\boldsymbol{\gamma} \in A_n$

$$(\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' J_n(\bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma} - \boldsymbol{\beta}^*) \geq e^{-3} (\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' J_n(\boldsymbol{\beta}^*)(\boldsymbol{\gamma} - \boldsymbol{\beta}^*). \tag{20}$$

Let $\tau = \exp(-3)/2$. Using (20), the probability in (19) can be bounded from above by

$$P[\sup_{\boldsymbol{\gamma}\in\partial A_n}(\boldsymbol{\gamma}-\boldsymbol{\beta})'s_n(\boldsymbol{\beta}) > \tau d_n^2\lambda_{\min}(J_n(\boldsymbol{\beta})) - \alpha_1 n d_n^2] \\ + P[\inf_{\boldsymbol{\gamma}\in\partial A_n}(\boldsymbol{\gamma}-\boldsymbol{\beta})'J_n(\bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma}-\boldsymbol{\beta})/2 < \tau d_n^2\lambda_{\min}(J_n(\boldsymbol{\beta}))]. \tag{21}$$

Let $\lambda_1^- = \lambda_{\min}(J(\boldsymbol{\beta}))/2$. Assuming $\alpha_1 < \lambda_1^-\tau$, the first probability in (21) can be bounded by

$$P[d_n||s_n(\boldsymbol{\beta})|| > \tau n d_n^2\lambda_1^- - \alpha_1 n d_n^2] + P[\lambda_{\min}(J_n(\boldsymbol{\beta})) < \lambda_1^- n] \\ \le P[||s_n(\boldsymbol{\beta})|| > (\tau\lambda_1^- - \alpha_1)n^{1/2}a_n^{1/2}] \\ + P[nd_n < n^{1/2}a_n^{1/2}] + P[\lambda_{\min}(J_n(\boldsymbol{\beta})) < \lambda_1^- n]. \tag{22}$$

Consider the first probability in (22). Note that $s_n(\boldsymbol{\beta}^*)$ is a random vector with zero mean and the covariance matrix $K_n(\boldsymbol{\beta}^*)$. Using Markov's inequality, the fact that $\mathrm{cov}[s_n(\boldsymbol{\beta}^*)] = nK(\boldsymbol{\beta}^*)$ and taking $\alpha_1 < \lambda^-\tau$ it can be bounded from above by

$$\frac{tr\{\mathrm{cov}[s_n(\boldsymbol{\beta}^*)]\}}{(\tau\lambda^- - \alpha_1)^2 n^2 d_n^2} = \frac{tr[K_n(\boldsymbol{\beta}^*)]}{(\tau\lambda^- - \alpha_1)^2 n^2 d_n^2} \le \frac{n\kappa p}{(\tau\lambda^- - \alpha_1)^2 n^2 d_n^2} \\ \le \frac{\kappa p}{(\tau\lambda^- - \alpha_1)^2 a_n} \to 0, \tag{23}$$

where the last convergence follows from $a_n \to \infty$.

The convergence to zero of the second probability in (22) follows from $nd_n^2/a_n \xrightarrow{P} \infty$. As eigenvalues of a matrix are continuous functions of its entries, we have $\lambda_{\min}(n^{-1}J_n(\boldsymbol{\beta}^*)) \xrightarrow{P} \lambda_{\min}(J(\boldsymbol{\beta}^*))$. Thus the convergence to zero of the third probability in (22) follows from the fact that in view of (A1) matrix $J(\boldsymbol{\beta}^*)$ is positive definite. The second term in (21) is bounded from above by

$$P[\inf_{\boldsymbol{\gamma}\in\partial A_n}(\boldsymbol{\gamma}-\boldsymbol{\beta})'J_n(\bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma}-\boldsymbol{\beta})/2 < \tau d_n^2\lambda_{\min}(J_n(\boldsymbol{\beta}))] \\ \le P[\inf_{\boldsymbol{\gamma}\in\partial A_n}(\boldsymbol{\gamma}-\boldsymbol{\beta})'[J_n(\bar{\boldsymbol{\gamma}}) - 2\tau J_n(\boldsymbol{\beta})](\boldsymbol{\gamma}-\boldsymbol{\beta})/2 \\ + 2\tau d_n^2\lambda_{\min}(J_n(\boldsymbol{\beta}))/2 < \tau d_n^2\lambda_{\min}(J_n(\boldsymbol{\beta}))] \\ \le P[\inf_{\boldsymbol{\gamma}\in\partial A_n}(\boldsymbol{\gamma}-\boldsymbol{\beta})'[J_n(\bar{\boldsymbol{\gamma}}) - 2\tau J_n(\boldsymbol{\beta})](\boldsymbol{\gamma}-\boldsymbol{\beta})/2 < 0] \to 0,$$

where the last convergence follows from Lemma 4 and (18).

Lemma 6 *Assume (A2) and (A3). Then we have* $\max_{i\le n}||\mathbf{x}_i||^2 a_n/n \xrightarrow{P} 0$.

Proof Using Markov inequality, (A2) and (A3) we have that $||\mathbf{x}_n||^2 a_n/n \xrightarrow{P} 0$. We show that this implies the conclusion. Denote $g_n := \max_{1\le i\le n}||\mathbf{x}_i||^2 a_n/n$ and $h_n := ||\mathbf{x}_n||^2 a_n/n$. Define sequence n_k such that $n_1 = 1$ and $n_{k+1} = \min\{n > n_k : \max_{i\le n}||\mathbf{x}_i||^2 > \max_{i\le n_k}||\mathbf{x}_i||^2\}$ (if such n_{k+1} does not exist put $n_{k+1} = n_k$). Without loss of generality we assume that for $A = \{n_k \to \infty\}$ we have $P(A) = 1$

as on A^c the conclusion is trivially satisfied. Observe that $g_{n_k} = h_{n_k}$ and $h_{n_k} \xrightarrow{P} 0$ as a subsequence of $h_n \xrightarrow{P} 0$ and thus also $g_{n_k} \xrightarrow{P} 0$. This implies that for any $\epsilon > 0$ there exists $n_0 \in \mathbf{N}$ such that for $n_k > n_0$ we have $P[|g_{n_k}| \leq \epsilon] \geq 1 - \epsilon$. As for $n \in (n_k, n_{k+1})$ $g_n \leq g_{n_k}$ since a_n/n is nonincreasing we have that if $n \geq n_0$ $P[|g_n| \leq \epsilon] \geq 1 - \epsilon$ i.e. $g_n \xrightarrow{P} 0$.

A.3 Proof of Proposition 1

Assume first that $\tilde{\boldsymbol{\beta}}^* = 0$ and note that this implies $p(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}}^*) = p(\beta_0) = C \in (0, 1)$. From (8) we have

$$P(y=1) = \mathbf{E}(y) = \mathbf{E}[\mathbf{E}(y|\tilde{\mathbf{x}})] = \mathbf{E}[q(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}})] = \mathbf{E}[p(\beta_0^* + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}}^*)] = C. \quad (24)$$

Using (24) and (7) we get

$$\begin{aligned}\mathbf{E}(\tilde{\mathbf{x}}y) &= \mathbf{E}\{\mathbf{E}[\tilde{\mathbf{x}}y|\tilde{\mathbf{x}}]\} = \mathbf{E}\{\tilde{\mathbf{x}}\mathbf{E}[y|\tilde{\mathbf{x}}]\} = \mathbf{E}[\tilde{\mathbf{x}}q(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}})] \\ &= \mathbf{E}[\tilde{\mathbf{x}}p(\beta_0^* + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}}^*)] = \mathbf{E}(\tilde{\mathbf{x}})C.\end{aligned} \quad (25)$$

From (24) we also have

$$\mathbf{E}(\tilde{\mathbf{x}}y) = \mathbf{E}\tilde{\mathbf{x}}I\{y=1\} = \mathbf{E}(\tilde{\mathbf{x}}|y=1)P(y=1) = \mathbf{E}(\tilde{\mathbf{x}}|y=1)C.$$

Comparing the last equation and right-side term in (25) we obtain $\mathbf{E}(\tilde{\mathbf{x}}|y=1) = E\tilde{\mathbf{x}} = \mathbf{E}(\tilde{\mathbf{x}}|y=0)$. Assume now $\mathbf{E}(\tilde{\mathbf{x}}|y=1) = \mathbf{E}(\tilde{\mathbf{x}}|y=0)$ which implies as before that that $\mathbf{E}(\tilde{\mathbf{x}}|y=1) = \mathbf{E}(\tilde{\mathbf{x}})$. Thus

$$\mathbf{E}(\tilde{\mathbf{x}}y) = \mathbf{E}(\tilde{\mathbf{x}}|y=1)\mathbf{E}(y) = \mathbf{E}(\tilde{\mathbf{x}})\mathbf{E}(y). \quad (26)$$

Since $(\beta_0^*, \tilde{\boldsymbol{\beta}}^*)$ is unique it suffices to show that (7) and (8) are satisfied for $\tilde{\boldsymbol{\beta}}^* = 0$ and β_0^* such that $Ep(\beta_0^*) = P(Y=1)$. This easily follows from (26).

References

1. Bache K, Lichman M (2013) UCI machine learning repository. University of California, Irvine
2. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
3. Bogdan M, Doerge R, Ghosh J (2004) Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. Genetics 167:989–999
4. Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analitycal extensions. Psychometrika 52:345–370
5. Burnham K, Anderson D (2002) Model selection and multimodel inference. A practical information-theoretic approach. Springer, New York

6. Carroll R, Pederson S (1993) On robustness in the logistic regression model. J R Stat Soc B 55:693–706
7. Casella G, Giron J, Martinez M, Moreno E (2009) Consistency of Bayes procedures for variable selection. Ann Stat 37:1207–1228
8. Chen J, Chen Z (2008) Extended Bayesian Information Criteria for model selection with large model spaces. Biometrika 95:759–771
9. Chen J, Chen Z (2012) Extended BIC for small-n-large-p sparse glm. Statistica Sinica 22: 555–574
10. Claeskens G, Hjort N (2008) Model selection and model averaging. Cambridge University Press, Cambridge
11. Czado C, Santner T (1992) The effect of link misspecification on binary regression inference. J Stat Plann Infer 33:213–231
12. Fahrmeir L (1987) Asymptotic testing theory for generalized linear models. Statistics 1:65–76
13. Fahrmeir L (1990) Maximum likelihood estimation in misspecified generalized linear models. Statistics 4:487–502
14. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Stat Assoc 96:1348–1360
15. Foster D, George E (1994) The risk inflation criterion for multiple regression. Ann Stat 22: 1947–1975
16. Hjort N, Pollard D (1993) Asymptotics for minimisers of convex processes. Unpublished manuscript
17. Konishi S, Kitagawa G (2008) Information criteria and statistical modeling. Springer, New York
18. Lehmann E (1959) Testing statistical hypotheses. Wiley, New York
19. Li K, Duan N (1991) Slicing regression: a link-free regression method. Ann Stat 19(2):505–530
20. Qian G, Field C (2002) Law of iterated logarithm and consistent model selection criterion in logistic regression. Stat Probab Lett 56:101–112
21. Ruud P (1983) Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. Econometrica 51(1):225–228
22. Sin C, White H (1996) Information criteria for selecting possibly misspecified parametric models. J Econometrics 71:207–225
23. Zak-Szatkowska M, Bogdan M (2011) Modified versions of Baysian Information Criterion for sparse generalized linear models. Comput Stat Data Anal 5:2908–2924
24. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc B 67(2):301–320

Semiparametric Inference in Identification of Block-Oriented Systems

Mirosław Pawlak

Abstract In this paper, we give the semiparametric statistics perspective on the problem of identification of a class of nonlinear dynamic systems. We present a framework for identification of the so-called block-oriented systems that can be represented by finite-dimensional parameters and an infinite-dimensional set of nonlinear characteristics that run typically through a nonparametric class of univariate functions. We consider systems which are expressible exactly in this form and the case when they are approximative models. In the latter case, we derive projections, that is, solutions which minimize the mean L_2 error. The chief benefit of such an approach is to make classical nonparametric estimates amenable to the incorporation of constraints and able to overcome the celebrated curse of dimensionality and system complexity. The developed methodology is explained by examining semiparametric versions of popular block-oriented structures, i.e., Hammerstein, Wiener, and parallel systems.

1 Introduction

The aim of system identification is to build a mathematical model of a class of dynamic systems from observed input-output data. This is a well-examined subject in the field of theoretical and applied automatic control, signal processing as well as process dynamics. The comprehensive overview of classical and modern system identification theory and its applications can be found in [2, 17, 21, 31]. There, the common approach to identification of dynamic systems is based on parametric models and the corresponding identification algorithms rely on the classical theory of maximum likelihood and prediction error. In contrast, modern statistics, machine learning and econometrics offer a large class of estimation techniques where parametric

M. Pawlak (✉)
Department of Electrical and Computer Engineering, University of Manitoba,
Winnipeg, MB R3T5V6, Canada
e-mail: Miroslaw.Pawlak@umanitoba.ca

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_16

assumption is not required [6, 9, 15, 16, 18–20, 30]. This includes strategies ranging from fully nonparametric algorithms to a more restrictive semiparametric inference. Surprisingly until recently, there has been a little influence of this powerful methodology on system identification. In [29] the machine learning theory approach has been applied to the parametric identification problem for linear systems with infinite memory having the fading memory property. In [26] an overview of recent contributions on the use of the regularization and RKHS (Reproducing Kernel Hilbert Space) theories in the context of identification of linear systems is presented. In [13] the statistical nonparametric estimation theory has been fully utilized and extended to identification of large class of nonlinear dynamic systems. Various nonparametric methods have been examined including classical Nadaraya–Watson kernel estimates, estimates employing orthogonal series expansions and the nonparametric version of stochastic approximation. The extension of the nonparametric function estimation setting to system identification is often non-trivial as in this case the functional relationship between observed signals and the unknown characteristics is in the indirect form. This type of hidden information regarding the unknown functions yields often to the lack of uniqueness and requires solving some indirect (inverse) estimation problems. Furthermore, the observed data do not often meet the mixing dependence condition, the assumption commonly assumed in the nonparametric/semiparametric statistical inference.

In this paper we focus on a class of dynamic nonlinear systems, often referred to as block-oriented, that find numerous applications in control engineering, chemical dynamics and biomedical systems [11]. Block-oriented nonlinear systems are represented by a certain composition of linear dynamical and nonlinear static models. Hence, a block-oriented system is defined by the pair $(\lambda, \mathbf{m}(\bullet))$, where λ defines infinite-dimensional parameter representing impulse response sequences of linear dynamical subsystems, whereas $\mathbf{m}(\bullet)$ is a vector of nonparametric multidimensional functions describing nonlinear elements. In the parametric identification approach to block-oriented systems one assumes that both λ and $\mathbf{m}(\bullet)$ are known up to unknown finite dimensional parameters, i.e., $\lambda = \lambda(\vartheta)$ and $\mathbf{m}(\bullet) = \mathbf{m}(\bullet; \zeta)$ for ϑ, ζ being finite dimensional unknown parameters. There are numerous identification algorithms for estimating ϑ, ζ representing specific block-oriented systems, see [10, 11] for an overview of the subject. Although such methods are quite accurate, it is well known, however, that parametric models carry a risk of incorrect model specification. On the other hand, in the nonparametric setting λ and $\mathbf{m}(\bullet)$ are completely unspecified and therefore the corresponding nonparametric block-oriented system does not suffer from risk of misspecification. Nevertheless, since nonparametric estimation algorithms make virtually no assumptions about the form of $(\lambda, \mathbf{m}(\bullet))$ they tend to be slower to converge to the true characteristics of a block-oriented system than correctly specified parametric algorithms. Moreover, the convergence rate of nonparametric methods is inversely proportional to the dimensionality of input and interconnecting signals. This is commonly referred to as the “curse of dimensionality”. Nonparametric methods have attracted a great deal of attention in statistics and econometrics, see [20, 30] for an overview of the subject. The number of texts on nonparametric algorithms tailored to the needs of engineering and system identification in particular is much smaller, see [13, 19] for recent contributions.

In practice, we can accept intermediate models which lie between parametric and fully nonparametric cases. For this so called semiparametric models we specify a parametric form for some part of the model but we do not require the parametric assumption for the remaining parts of the model. Hence, the nonparametric description $(\lambda, \mathbf{m}(\bullet))$ of the system is now replaced by $(\theta, \mathbf{g}(\bullet))$, where θ is a finite dimensional vector and $\mathbf{g}(\bullet)$ is a set of nonparametric nonlinearities being typically univariate functions. The parameter θ represents characteristics of all linear dynamical subsystems and low-dimensional approximations of multivariate nonlinearities. The fundamental issue is to characterize the accuracy of approximation of $(\lambda, \mathbf{m}(\bullet))$ by the selected semiparametric model $(\theta, \mathbf{g}(\bullet))$. This challenging problem will be addressed in this paper in some specific cases. Once such characterization is resolved, we can make use of this low complexity semiparametric representation to design practical identification algorithms that share the efficiency of parametric modelling while preserving the high flexibility of the nonparametric case. In fact, in many situations we are able to identify linear and nonlinear parts of a block-oriented system under much weaker conditions on the system characteristics and underlying probability distributions.

A semiparametric inference is based on the concept of blending together parametric and nonparametric estimation methods. The basic idea is to first examine the parametric part of the block-oriented structure as if all nonparametric parts were known. To eliminate the dependence of a parametric fitting criterion on the characteristics of the nonparametric parts, we form pilot nonparametric estimates of the characteristics being indexed by a finite-dimensional vector of the admissible value of the parameter. Then, this is used to establish a parametric fitting criterion (such as least squares) with random functions representing all estimated nonparametric characteristics. The resulting parameter estimates are employed to form final nonparametric estimates of the nonlinear characteristics. As a result of this interchange, we need some data resampling schemes in order to achieve some statistical independence between the estimators of parametric and nonparametric parts of the system. This improves the efficiency of the estimates and facilitates the mathematical analysis immensely. In Sect. 2 we examine sufficient conditions for the convergence of our identification algorithms for a general class of semiparametric block-oriented systems. This general theory is illustrated (Sect. 3) in the case of semiparametric versions of Hammerstein systems, Wiener system, and parallel connections. We show in this context that the semiparametric strategy leads to consistent estimates of $(\theta, \mathbf{g}(\bullet))$ with optimal rates which are independent of the input signal dimensionality. These results are also verified in some simulation studies.

An overview of the theory and applications of semiparametric inference in statistics and econometrics can be found in [15, 20, 28, 34].

2 Nonparametric and Semiparametric Inference

The modern nonparametric inference provides a plethora of estimation methods allowing us to recover system characteristics with the minimum knowledge about their functional forms. This includes classical methods like $k-$nearest neighbours, kernel and series estimators. On the other hand, sparse basis functions, regularization techniques, support vector machines, and boosting methods define modern machine learning alternatives [16, 18, 23].

For a given set of training data $D_N = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_N, Y_N)\}$ taken at the input and output of a certain system, a generic nonparametric estimate of a regression function $m(\mathbf{x}) = E\{Y_t|\mathbf{X}_t = \mathbf{x}\}$ can be written as

$$\hat{m}_N(\mathbf{x}) = \sum_{t=1}^{N} \alpha_t Y_t \mathbf{K}(\mathbf{x}, \mathbf{X}_t), \tag{1}$$

where $\mathbf{K}(\mathbf{x}, \mathbf{v})$ is a kernel function and $\{\alpha_t\}$ is a weight sequence. For the classical kernel estimate $\{\alpha_t\}$ plays role of the normalizing sequence, i.e., $\alpha_t = (\sum_{i=1}^{N} \mathbf{K}(\mathbf{x}, \mathbf{X}_i))^{-1}$ for each t, where $\mathbf{K}(\mathbf{x}, \mathbf{v})$ is the kernel of the convolution type, i.e., $\mathbf{K}(\mathbf{x}, \mathbf{v}) = \mathbf{K}(\mathbf{x} - \mathbf{v})$. On the other hand, in support vector kernel machines, $\{\alpha_t\}$ is selected by the optimization algorithm defined by the maximal-margin separation principle and the kernel function is of the inner product type (Mercer's kernels) $\mathbf{K}(\mathbf{x}, \mathbf{v}) = \sum_l \phi_l(\mathbf{x})\phi_l(\mathbf{v})$. In order to achieve the desired consistency property, i.e., that $\hat{m}_N(\mathbf{x})$ tends to $m(\mathbf{x})$ as $N \to \infty$, the kernel function must be tuned locally. This can be achieved by introducing the concept of smoothing parameters that control the size of local information that is employed in the estimation process. For instance, in the kernel estimate we use $\mathbf{K}_b(\mathbf{x} - \mathbf{v}) = b^{-d}\mathbf{K}((\mathbf{x} - \mathbf{v})/b)$, for $\mathbf{x}, \mathbf{v} \in R^d$, where b is a positive number, usually called the bandwidth.

The consistency is the desirable property and is met by most classical nonparametric techniques. Some modern techniques like support vector machines are not local since they use the entire training data in the design process. This can be partially overcome by using the properly regularized kernels.

A serious limitation in the use of nonparametric estimators is the fact that they are prone to the dimensionality of observed signals as well as the smoothness of estimated characteristics [9, 20, 30]. To illustrate this point, let us consider the following multiple-input, single-output (MISO), nonlinear autoregressive model of order p:

$$Y_n = m(Y_{n-1}, Y_{n-2}, \ldots, Y_{n-p}, \mathbf{U}_n) + Z_n, \tag{2}$$

where $\mathbf{U}_n \in R^d$ is the input signal, Z_n is noise process, and $m(\bullet, \bullet)$ is a $(p + d)$—dimensional function. It is clear that $m(\bullet, \bullet)$ is a regression function of Y_n on the past outputs $Y_{n-1}, Y_{n-2}, \ldots, Y_{n-p}$ and the current input $\mathbf{U}_n$. Thus, it is a straightforward task to form a multivariate nonparametric regression estimate such as the one in (1), where the signal $\mathbf{X}_t \in R^{p+d}$ is defined as $\mathbf{X}_t = (Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}, \mathbf{U}_t)$. The convergence analysis, [9, 20], of such an estimate will strongly depend on the stability conditions of the nonlinear recursive difference equation:

$$y_n = m(y_{n-1}, y_{n-2}, \ldots, y_{n-p}, \mathbf{u}_n).$$

With this respect, a fading-memory type assumption along with the Lipschitz continuity of $m(\bullet, \bullet)$ seem to be sufficient for the consistency of nonparametric regression estimates. Hence, for $m(\bullet, \bullet)$ being a Lipschitz continuous function the best possible rate can be $O_P\left(N^{-\frac{1}{2+p+d}}\right)$, where $O_P(\bullet)$ denotes the convergence in probability. For the second order system $p = 2$ and two-dimensional input this gives a very slow rate of $O_P(N^{-1/6})$. This apparent curse of dimensionality also exists in the case of the MISO Hammerstein system which will be examined in the next section.

To overcome this problem one can consider to approximate the regression function $m(\mathbf{x}) = E\{Y_t|\mathbf{X}_t = \mathbf{x}\}$, $\mathbf{x} \in R^q$, by some low-dimensional structures. We are looking for a parsimonious semiparametric alternative which can be represented by a finite-dimensional parameter and a set of single-variable nonlinearities. The following is a simple semiparametric model for $m(\mathbf{x})$:

$$\mu(\mathbf{x}) = g(\theta^T \mathbf{x}), \tag{3}$$

where the function $g(\bullet)$ and the parameter $\theta \in R^q$ are unknown and must be estimated. We note that $g(\bullet)$ is a single variable function and thus the curse of dimensionality for the model $\mu(\mathbf{x})$ is removed. The model $\mu(\mathbf{x})$ is not uniquely defined. In fact if $g(\bullet)$ is linear then we cannot identify θ. Moreover, the scaling of the vector θ does not influence the values of $g(\bullet)$ if we rescale $g(\bullet)$ accordingly. Hence, we need to normalize θ either by setting one of the coefficients of θ to one, e.g., $\theta_1 = 1$ or by imposing the restriction $||\theta|| = 1$. We will call the set of all such normalized values of θ as Θ.

In system identification the vector of covariates $\mathbf{X}_t$ can be decomposed as $\mathbf{X}_t = (\mathbf{U}_t, \mathbf{V}_t)$, where $\mathbf{U}_t \in R^d$ has the interpretation of the input signal and $\mathbf{V}_t \in R^p$ defines the past values of the output signal. Then we can propose a few alternatives to the model in (3), e.g.,

$$\mu(\mathbf{u}, \mathbf{v}) = \rho^T \mathbf{v} + g(\gamma^T \mathbf{u}), \tag{4}$$

where $\rho \in R^p$ and $\gamma \in R^d$ are unknown parameters. This semiparametric model applied in (2) would result in a partially linear nonlinear system of the form

$$Y_n = \rho_1 Y_{n-1} + \rho_2 Y_{n-2} + \cdots + \rho_p Y_{n-p} + g(\gamma^T \mathbf{U}_n) + Z_n. \tag{5}$$

The statistical inference for the model in (3), i.e., estimation of $(\theta, g(\bullet))$ requires the characterization of the "true" characteristics $(\theta^*, g^*(\bullet))$. This can be done by finding the optimal L_2 projection of the original system onto the model defined in (3). Hence, we wish to minimize the risk function

$$Q(\theta, g(\bullet)) = E\{(Y_t - g(\theta^T \mathbf{X}_t))^2\} \tag{6}$$

with respect to $\theta \in \Theta$ and $g(\bullet)$ such that $E\{g^2(\theta^T \mathbf{X}_t))\} < \infty$. The minimizer of $Q(\theta, g(\bullet))$ will be denoted as $(\theta^*, g^*(\bullet))$. Since the minimization of $Q(\theta, g(\bullet))$ is equivalent to the minimization of $E\{(Y_t - g(\theta^T \mathbf{X}_t))^2 | \mathbf{X}_t\}$. This is the L_2 projection and for a given $\theta \in \Theta$ the solution is $g(w; \theta) = E\{Y_t | \theta^T \mathbf{X}_t = w\}$. This is just a regression function of Y_t on $\theta^T \mathbf{X}_t$, i.e., the best predictor of the output signal Y_t by the projection of $\mathbf{X}_t$ onto the direction defined by the vector θ. Plugging this choice into $Q(\theta, g(\bullet))$, i.e., calculating $Q(\theta, g(\bullet; \theta))$ we can readily obtain the following score function

$$Q(\theta) = E\{(Y_t - g(\theta^T \mathbf{X}_t; \theta)^2\} = E\{(var(Y_t | \theta^T \mathbf{X}_t)\}. \tag{7}$$

The minimizer of $Q(\theta)$ with respect to $\theta \in \Theta$ defines the optimal θ^* and consequently the corresponding optimal nonlinearity $g^*(w) = g(w; \theta^*)$. This yields the minimal risk $Q^*_{sp} = Q(\theta^*)$ for the pre-selected semiparametric model. It is clear that this risk is larger than the Bayes risk Q^* corresponding to the fully nonparametric approach. On the other hand, Q^*_{sp} is smaller than the risk Q^*_p corresponding to the parametric choice. Hence, we have the following relationship between the aforementioned modelling strategies

$$Q^* \leq Q^*_{sp} \leq Q^*_p.$$

It should be noted, however, that these inequalities only hold for the limit case, i.e., if the training sample size is infinite. In a practical finite sample size situation, there is no preferable modelling approach that gives the smallest generalized identification error.

In the semiparametric case characterized by (7), it is difficult to determine an explicit formula for the projection $g(w; \theta)$ and then to find θ^* minimizing the risk in (7). It is clear that the smoothness and shape of $g(w; \theta)$ is controlled by the smoothness of $m(\mathbf{x})$ and the conditional distribution of $\mathbf{X}_t$ given $\theta^T \mathbf{X}_t$. To shed some light on this issue let us consider a simple example.

Example 1. Let $Y_t = m(U_t, U_{t-1}) + Z_t$ be the nonlinear autoregressive regression model with $m(x_1, x_2) = x_1 x_2$. Assume that $\{U_t\}$ is zero mean unit variance stationary Gaussian process with the correlation function $E\{U_{t+\tau} U_t\} = \rho(\tau)$. Let us also denote $\rho = \rho(1)$. The noise process $\{Z_t\}$ is assumed to be a stationary process with zero mean and unit variance being, moreover, independent of $\{U_t\}$. We wish to determine the best L_2 approximation of the system $Y_t = U_t U_{t-1} + Z_t$ by a semiparametric model of the form $g(U_t + \theta U_{t-1})$. The aforementioned discussion reveals that first we have to determine the projection $g(w; \theta) = E\{Y_t | U_t + \theta U_{t-1} = w\}$ and next to find the optimal θ^* by minimizing the risk $Q(\theta) = E\{var(Y_t | U_t + \theta U_{t-1})\}$. To do so, let us first note that the random vector $(U_{t-1}, U_t + \theta U_{t-1})$ has the bivariate Gaussian distribution with zero mean and covariance matrix

$$\begin{pmatrix} 1 & \rho + \theta \\ \rho + \theta & 1 + \theta^2 + 2\theta\rho \end{pmatrix}.$$

This fact and some algebra yield

$$g(w;\theta) = a(\theta)w^2 + b(\theta), \tag{8}$$

where $a(\theta) = (\rho + \theta(1-\theta))/(1+\theta^2+2\theta\rho)$ and $b(\theta) = -\theta(1-\rho^2))/(1+\theta^2+2\theta\rho)$. Further algebra leads also to an explicit formula for the projection error $Q(\theta)$ for a given θ. In Fig. 1 we depict $Q(\theta)$ as a function of θ for the value of the correlation coefficient ρ equal to 0, 0.4, 0.8. The dependence of $Q(\theta)$ on negative values of ρ is just a mirror reflection of the curves shown in Fig. 1. Interestingly, we observe that in the case of $\rho = 0$ we have two values of θ minimizing $Q(\theta)$, i.e., $\theta^* = \pm 1/\sqrt{3}$. When $|\rho|$ is increasing, the optimal θ is unique and is slowly decreasing from $\theta^* = 0.577$ for $\rho = 0$ to $\theta^* = 0.505$ for $\rho = 0.9$. On the hand, the value of the minimal error $Q(\theta^*)$ is decreasing fast from $Q(\theta^*) = 0.75$ for $\rho = 0$ to $Q(\theta^*) = 0.067$ for $\rho = 0.9$. Figure 2 shows the optimal nonlinearities $g^*(w) = g(w;\theta^*)$ corresponding to the values $\rho = 0, 0.4, 0.8$. It is worth noting that $g^*(w)$ for $\rho = 0$ is smaller than $g^*(w)$ for any $\rho > 0$. Similar relationships hold for $\rho < 0$. Thus, we can conclude that the best approximation (for $\rho = 0$) of the system $Y_t = U_tU_{t-1} + Z_t$ by the class of semiparametric models $\{g(U_t + \theta U_{t-1}) : \theta \in \Theta\}$ is the model $Y_t = g^*(U_t + \theta^* U_{t-1}) + Z_t$, where $\theta^* = \pm 1/\sqrt{3}$ and $g^*(w) = \pm\frac{\sqrt{3}-1}{4}w^2 \mp \frac{\sqrt{3}}{4}$. In the case when, e.g., $\rho = 0.5$ we obtain $\theta^* = 0.532$ and $g^*(w) = 0.412w^2 - 0.219$. We should also observe that our semiparametric model represents the Wiener cascade system with the impulse response $(1, \theta^*)$ and the nonlinearity $g^*(w)$, see Sect. 3 for further discussion on Wiener systems. The fact that the correlation reduces the value of the projection error $Q(\theta)$ can be interpreted as follows. With an increasing correlation between input variables the bivariate function $m(U_t, U_{t-1}) = U_tU_{t-1}$ behaves like a function of a single variable. In fact, from (8) we have that $b(\theta) \to 0$ as $|\rho| \to 1$.

Thus far, we have discussed the preliminary aspects of the semiparametric inference concerning the characterization of the optimal characteristics $(\theta^*, g^*(\bullet))$ of the model in (3). Next, we wish to set up estimators of θ^* and $g^*(w)$. If the regression

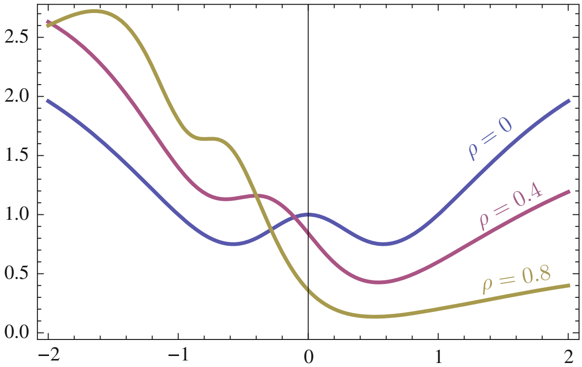


Fig. 1 The projection error $Q(\theta)$ versus θ for the values of the correlation coefficient $\rho = 0, 0.4, 0.8$

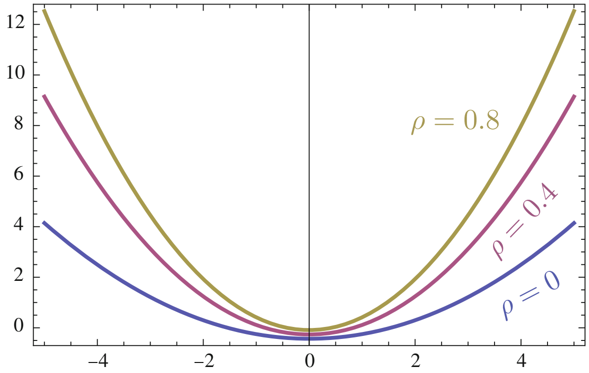


Fig. 2 The optimal nonlinearity $g^{\star}(w)$ for the values of the correlation coefficient $\rho = 0, 0.4, 0.8$

function $g(w; \theta) = E\{Y_t | \theta^T \mathbf{X}_t = w\}$, $\theta \in \Theta$ is assumed to be known, then, due to (7), an obvious estimator of θ^* would be a minimizer of the following empirical counterpart of $Q(\theta)$:

$$Q_N(\theta) = N^{-1} \sum_{t=1}^{N} (Y_t - g(\theta^T \mathbf{X}_t; \theta))^2. \tag{9}$$

Since $g(w; \theta)$ is unknown, this is not a feasible estimator. We can, however, estimate the regression function $g(w; \theta)$ by some standard nonparametric methods like kernel algorithms, see (1). Let $\hat{g}(w; \theta)$ denote a nonparametric estimate of $g(w; \theta)$. As a concrete choice we use the classical kernel estimate

$$\hat{g}(w; \theta) = \sum_{t=1}^{N} Y_t K((w - \theta^T \mathbf{X}_t)/b) / \sum_{l=1}^{N} K((w - \theta^T \mathbf{X}_l)/b), \tag{10}$$

where b is the bandwidth parameter that controls the accuracy of the estimate.

In the limit any reasonable nonparametric estimate $\hat{g}(w; \theta)$ is expected to tend to $g(w; \theta)$ which, in turn, satisfies the restriction $g(w; \theta^*) = g^*(w)$. Hence, substituting $g(w; \theta)$ in (9) by $\hat{g}(w; \theta)$ we can obtain the following criterion depending solely on θ:

$$\hat{Q}_N(\theta) = N^{-1} \sum_{t=1}^{N} (Y_t - \hat{g}(\theta^T \mathbf{X}_t; \theta))^2. \tag{11}$$

It is now natural to define an estimate $\hat{\theta}$ of θ^* as the minimizer of $\hat{Q}_N(\theta)$, i.e.,

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \hat{Q}_N(\theta). \tag{12}$$

This approach may lead to an effective estimator of θ^* subject to some limitations. First, as we have already noted, we should be able to estimate the projection $g(w; \theta)$

for a given θ. In the context of block-oriented systems, the difficulty of this step depends on the complexity of the studied nonlinear system, i.e., whether nonlinear components can be easily estimated as if the parametric part of the system were known. This is due to the fact that some intermediate signals in block-oriented systems are not measured. Nevertheless, in the next section we will demonstrate that effective identification algorithms are feasible for semiparametric versions of Hammerstein, Wiener and parallel systems. Second, we must minimize the criterion $\hat{Q}_N(\theta)$ which may be an expensive task mostly if θ is highly dimensional and if the gradient vector of $\hat{Q}_N(\theta)$ is difficult to evaluate. To partially overcome these computational difficulties we can use the following generic iterative algorithm:

Step 1: Select an initial $\hat{\theta}^{(\text{old})}$ and set $\hat{g}(w; \hat{\theta}^{(\text{old})})$.
Step 2: Minimize the criterion

$$\widetilde{Q}_N(\theta) = N^{-1} \sum_{t=1}^{N} (Y_t - \hat{g}(\theta^T \mathbf{X}_t; \theta^{(\text{old})}))^2 \tag{13}$$

with respect to θ and use the obtained value $\hat{\theta}^{(\text{new})}$ to update $\hat{g}(w; \theta)$, i.e., go to *Step* 1 in order to get $\hat{g}(w; \theta^{(\text{new})})$.
Step 3: Iterate between the above two steps until a certain stopping rule is satisfied.

Note that in the above algorithm the criterion $\widetilde{Q}_N(\theta)$ has a weaker dependence on θ than the original criterion $\hat{Q}_N(\theta)$. In fact, in $\widetilde{Q}_N(\theta)$ the estimate $\hat{g}(w; \theta)$ of the projection $g(w; \theta)$ is already specified.

Concerning the recovery of the model optimal nonlinearity $g^*(\bullet)$ we can use the estimate $\hat{\theta}$ and plug it back into our pilot estimate $\hat{g}(w; \theta)$ to obtain $\hat{g}(w; \hat{\theta})$. This can define a nonparametric estimate of $g^*(\bullet)$. Nevertheless, one can use any other nonparametric estimate $\tilde{g}(\bullet; \theta)$ with θ replaced by $\hat{\theta}$. Yet another important issue is the problem of selecting a smoothing parameter, like the bandwidth b in the kernel estimate in (10), which tunes nonparametric estimates $\hat{g}(\bullet; \theta)$ and $\hat{g}(\bullet)$. The former estimate is used as a preliminary estimator of the projection $g(\bullet; \theta)$ so that θ^* can be estimated in the process of minimizing $\hat{Q}_N(\theta)$ in (11). On the other hand, the latter estimate is used as a final estimate of $g^*(\bullet)$. Hence, we may be forced to select two separate smoothing parameters. The one for $\hat{g}(\bullet; \theta)$, and the other for $\hat{g}(\bullet)$. This can be done by augmenting the definition of $\hat{Q}_N(\theta)$ in (11) by adding the smoothing parameter as a variable in $\hat{Q}_N(\theta)$. Hence, we define $\hat{Q}_N(\theta, b)$ and then minimize $\hat{Q}_N(\theta, b)$ simultaneously with respect to θ and b. The bandwidth obtained in this process is by no means good for selecting the estimate $\hat{\theta}$. Whether this is the proper choice for the accurate estimation of $\hat{g}(\bullet)$ is not quite clear, see [14] for the affirmative answer to this controversy in the context of the classical regression problem from $i.i.d.$ data.

In order to establish consistency properties of the aforementioned estimates we first need to establish that the criterion $\hat{Q}_N(\theta)$ in (11) tends (P) as $N \to \infty$ to the limit criterion $Q(\theta)$ in (7) for a given $\theta \in \Theta$. This holds under fairly general conditions due to the law of large numbers. Furthermore, as we have already argued we identify the optimal θ^* with the minimum of $Q(\theta)$. Note, however, that $\hat{Q}_N(\theta)$ is not a convex function of θ and therefore need not achieve a unique minimum. This, however, is of no serious importance for the consistency since we may weaken our requirement on the minimizer $\hat{\theta}$ of $\hat{Q}_N(\theta)$ and define $\hat{\theta}$ as any estimator that nearly minimizes $\hat{Q}_N(\theta)$, i.e.,

$$\hat{Q}_N(\hat{\theta}) \le \inf_{\theta \in \Theta} \hat{Q}_N(\theta) + \delta_N, \tag{14}$$

for any random sequence δ_N, such that $\delta_N \to 0(P)$, see [28]. It is clear that (14) implies that $\hat{Q}_N(\hat{\theta}) \le \hat{Q}_N(\theta^*) + \delta_N$ and this is sufficient for the convergence of $\hat{\theta}$ defined in (14) to θ^*.

Since $\hat{\theta}$ depends on the whole mapping $\theta \mapsto \hat{Q}_N(\theta)$, the convergence of $\hat{\theta}$ to θ^* requires uniform consistency of the corresponding criterion function, i.e., we need $\sup_{\theta \in \Theta} \left| \hat{Q}_N(\theta) - Q(\theta) \right| \to 0(P)$. This uniform convergence is the essential step in proving the convergence of $\hat{\theta}$ to θ^*. This can be established by using formal techniques for verifying whether the following class of functions

$$\{(y - \hat{g}(w; \theta))^2 : \theta \in \Theta\}$$

satisfies a uniform law of large numbers that is often referred to as the Glivienko-Cantelli property [28]. This along with the assumption that the limit criterion $Q(\theta)$ is a continuous function on the compact set $\Theta \subset R^q$, such that $\theta^* \in \Theta$, imply that for any sequence of estimators $\hat{\theta}$ that satisfy (14) we have

$$\hat{\theta} \to \theta^* \quad \text{as} \quad N \to \infty(P).$$

A related issue of interest pertaining to a given estimate is the rate of convergence, i.e., how fast the estimate tends to the true characteristic. Under the differentiability condition of the mapping $\theta \mapsto (\bullet - \hat{g}(\bullet; \theta))^2$ we can consider the problem of the convergence rate. Hence, if $Q(\theta)$ admits the second-order Taylor expansion at $\theta = \theta^*$ then for $\hat{\theta}$ defined in (14) with $N\delta_N \to 0(P)$, we have

$$\hat{\theta} = \theta^* + O_P(N^{-1/2}). \tag{15}$$

This is the usual parametric rate of convergence.

Since $\hat{\theta} \to \theta^*$ then it is reasonable to expect that the estimate $g(\bullet; \hat{\theta})$ converges to $g(\bullet; \theta^*) = g^*(\bullet)$. The following decomposition will facilitate this claim

$$\begin{aligned} g(\bullet; \hat{\theta}) - g^*(\bullet) = \{&\hat{g}(\bullet; \hat{\theta}) - \hat{g}(\bullet; \theta^*)\} \\ &+ \{\hat{g}(\bullet; \theta^*) - g^*(\bullet)\}. \end{aligned} \tag{16}$$

The convergence (P) of the second term to zero in the above decomposition represents a classical problem in nonparametric estimation. The difficulty of establishing this convergence depends on the nature of the dependence between random signals within the underlying system. Concerning the first term in (16) we can apply the linearization technique, i.e.,

$$\hat{g}(\bullet;\hat{\theta}) - \hat{g}(\bullet;\theta^*) = \left\{ \frac{\partial}{\partial\theta}\hat{g}(\bullet;\theta)_{|\theta=\theta^*} \right\}^T (\hat{\theta}-\theta^*) \\ + o_P(\|\hat{\theta}-\theta^*\|).$$

To show the convergence (P) of the first term to zero it suffices to prove that the derivative has a finite limit (P). This fact can be directly verified for a specific estimate $\hat{g}(\bullet;\theta)$ of $g(\bullet;\theta)$. Hence, the statistical accuracy of $\hat{g}(\bullet;\hat{\theta})$ is determined by the second term of the decomposition in (16). It is well known that the common nonparametric estimates reveal the rate $\hat{g}(\bullet;\theta^*) = g^*(\bullet) + O_P(N^{-\beta})$, where $1/3 \leq \beta < 1$ depends on the smoothness of $g^*(\bullet)$, see [13, 20, 30]. Thus, we obtain

$$\hat{g}(\bullet;\hat{\theta}) = g^*(\bullet) + O_P(N^{-\beta}). \tag{17}$$

For instance, if $g^*(\bullet)$ is the Lipschitz continuous function and $\hat{g}(\bullet;\hat{\theta})$ is the kernel nonparametric estimate then (17) holds with $\beta = 1/3$. For twice differentiable nonlinearities, this takes place with $\beta = 2/5$.

The criterion $\hat{Q}_N(\theta)$ in (11) utilizes the same data to form the pilot nonparametric estimate $\hat{g}(w;\theta)$ and to define $\hat{Q}_N(\theta)$. This is not generally a good strategy and some form of resampling scheme should be applied in order to separate measurements into the testing data (used to form $\hat{Q}_N(\theta)$) and training sequence (used for forming the estimate $\hat{g}(w;\theta)$). Such separation will facilitate not only the mathematical analysis of the estimation algorithms but also gives a desirable separation of parametric and nonparametric estimation problems, which allows one to evaluate parametric and nonparametric estimates more precisely. One such a strategy would be the leave-one-out method which modifies $\hat{Q}_N(\theta)$ as follows

$$\bar{Q}_N(\theta) = N^{-1}\sum_{t=1}^{N}(Y_t - \hat{g}_t(\theta^T\mathbf{X}_t;\theta))^2, \tag{18}$$

where $\hat{g}_t(w;\theta)$ is the version of the estimate $\hat{g}(w;\theta)$ with the training data pair $(\mathbf{X}_t, Y_t)$ omitted from calculation. For instance, in the case of the kernel estimate in (10) this takes the form

$$\hat{g}_t(w;\theta) = \sum_{i\neq t}^{N} Y_i K((w-\theta^T\mathbf{X}_i)/b) \Big/ \sum_{l\neq t}^{N} K((w-\theta^T\mathbf{X}_l)/b).$$

Yet another efficient resampling scheme is based on the partition strategy which reorganizes a set of training data D_N into two non overlapping subsets that are dependent as weakly as possible. Hence, we define two non overlapping subsets T_1, T_2 of the training set D_N such that T_1 is used to estimate the projection $g(w;\theta)$

whereas T_2 is used as a testing sequence to form the least-squares criterion $\hat{Q}_N(\theta)$ in (11). There are various strategies to split the data for the efficient estimation of θ^* and $g^*(\bullet)$. The machine learning principle says the testing sequence T_2 should consist (if it is feasible) of independent observations, whereas the training sequence T_1 can be quite arbitrary [6].

3 Semiparametric Block-Oriented Systems

In this section, we will illustrate the semiparametric methodology developed in Sect. 2 by examining a few concrete cases of block-oriented systems. This includes popular Hammerstein, Wiener, and parallel structures.

3.1 Semiparametric Hammerstein Systems

Let us begin with the multiple-input, single-output (MISO) Hammerstein system which is depicted in in Figure 3. This is a cascade system consisting of a nonlinear characteristic $m(\cdot)$ followed by a linear dynamic subsystem with the impulse response function $\{\lambda_n\}$. Only the input $\boldsymbol{U}_n$ and output Y_n signals are measured, the other intermediate signals are not available. The fully nonparametric Hammerstein system is given by the following input-output relationship:

$$\begin{aligned} Y_n &= \Lambda(z)V_n + H(z)e_n, \\ V_n &= m(\mathbf{U}_n), \end{aligned} \tag{19}$$

where $\Lambda(z)$ is a causal transfer function defined as $\Lambda(z) = \sum_{i=0}^{\infty} \lambda_i z^{-i}$, with z^{-1} being the backward-shift operator, i.e., $\Lambda(z)V_n = \sum_{i=0}^{\infty} \lambda_i V_{n-i}$.

Moreover, $\sum_{i=0}^{\infty} |\lambda_i| < \infty$ and $\Lambda(z)$ is stable, i.e., $\Lambda(z)$ has poles within the unit circle. The noise transfer function, $H(z) = \sum_{i=0}^{\infty} h_i z^{-i}$ satisfies $\sum_{i=0}^{\infty} |h_i| < \infty$ and has both poles and zeros within the unit circle. The latter requirement is imposing the assumption that the noise model is stable and inversely stable, see [21]. The input noise process $\{e_n\}$ is white with a finite variance.

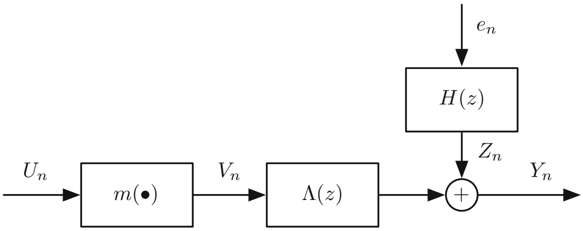


Fig. 3 MISO Hammerstein system

The system in in Fig. 3 is excited by the d-dimensional input $\mathbf{U}_n$, which is assumed to be a sequence of $i.i.d.$ random vectors. The output of the linear dynamic subsystem is corrupted by an additive noise Z_n being independent of $\{\mathbf{U}_n\}$. The system nonlinearity $m(\bullet)$ is a nonparametric measurable function defined on R^d such that $E|m(\mathbf{U}_n)| < \infty$.

It is known, see [13], that if $\Lambda(\infty) = 1$ and $E\{m(\mathbf{U}_n)\} = 0$ then $m(\mathbf{u}) = E\{Y_n|\mathbf{U}_n = \mathbf{u}\}$. This fact holds for any correlated noise process. This key identity allows us to recover $m(\bullet)$ by applying nonparametric regression estimates such those defined in (1). The technical obstacles include the fact that the internal signal V_n is not measured and that the output process $\{Y_n\}$ need not be of mixing type. In fact, let $Y_n = 0.5Y_{n-1} + m(U_n)$ be the Hammerstein system with the $AR(1)$ linear process and the nonlinearity being $m(u) = 1$ for $u \geq 0$ and $m(u) = 0$ otherwise. Then, using the result proved in [1] we can conclude that $\{Y_n\}$ is not strongly-mixing.

Let $\hat{m}_N(\bullet)$ be a nonparametric regression function estimate based on the training data $D_N = \{(\mathbf{U}_1, Y_1), \ldots, (\mathbf{U}_N, Y_N)\}$. It can be demonstrated (under common smoothing conditions on $m(\mathbf{u})$) that for a large class of nonparametric regression estimates, see [13], we have

$$\hat{m}_N(\mathbf{u}) = m(\mathbf{u}) + O_P\left(N^{-2/(d+4)}\right). \tag{20}$$

Hence, the estimates suffer the curse of dimensionality since the rate of convergence gets slower as d increases. It is also worth noting that the linear part of the system $\Lambda(z)$ can be recovered via the correlation method independently on the form of the system nonlinearity and the noise structure, see [13]. This defines a fully nonparametric identification strategy for the MISO Hammerstein system. The statistical accuracy, however, of such estimation algorithms is rather low due to the generality of the problem.

In many practical situations and due to the inherent complexity of the nonparametric Hammerstein system it is sufficient if we resort to the following semiparametric alternative of (19) (see Fig. 4)

$$\begin{aligned} Y_n &= \Lambda(z;\lambda)V_n + H(z;\lambda)e_n, \\ V_n &= g(\gamma^T\mathbf{U}_n), \end{aligned} \tag{21}$$

where $\Lambda(z;\lambda)$ and $H(z;\lambda)$ are parametrized rational transfer functions. The function $g(\bullet)$ and the d—dimensional parameter γ define the one-dimensional semiparamet-

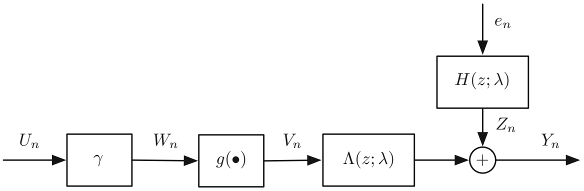


Fig. 4 Semiparametric MISO Hammerstein model

ric approximation of $m(\bullet)$ which was already introduced in Sect. 2, see (3). Note the class of dynamical systems represented by the rational transfer functions covers a wide range of linear autoregressive and moving average processes.

Hence, the semiparametric model in (21) is characterised by the pair $(\theta, g(\bullet))$, where $\theta = (\lambda, \gamma)$. Since the identifiability of the model requires that $\Lambda(\infty; \lambda) = 1$ and $\gamma_1 = 1$, therefore we can define the parameter space as $\Theta = \{(\lambda, \gamma) : \Lambda(\infty; \lambda) = 1, \gamma_1 = 1\}$, such that Θ is a compact subset of R^{p+d}, where p is the dimensionality of λ. In order to develop constructive identification algorithms let us define the concept of the true Hammerstein system corresponding to (21). We may assume without loss of generality that the true system is in the form as in (21) and this will be denoted by the asterisk sign, i.e., the true system is defined as

$$\begin{aligned} Y_n &= \Lambda(z; \lambda^*) V_n + H(z; \lambda^*) e_n, \\ V_n &= g^*(\gamma^{*T} \mathbf{U}_n), \end{aligned} \tag{22}$$

where it is natural to expect that $\theta^* \in \Theta$.

Since the dependence of Y_n on V_n is linear then we can recall, see [21], that a one-step ahead prediction error for a given $\theta \in \Theta$ is given by

$$\varepsilon_n(\theta) = H^{-1}(z; \lambda) \left[Y_n - \Lambda(z; \lambda) V_n(\gamma) \right], \tag{23}$$

where $V_n(\gamma)$ is the counterpart of the true signal V_n corresponding to the value γ. Under our normalization we note that for a given $\gamma^T \mathbf{U}_n$ the best L_2 predictor of $V_n(\gamma)$ is the regression $E\{V_n(\gamma) | \gamma^T \mathbf{U}_n\} = E\{Y_n | \gamma^T \mathbf{U}_n\}$. Hence, let $g(w; \gamma) = E\{Y_n | \gamma^T \mathbf{U}_n = w\}$ be the regression function predicting the unobserved signal $V_n(\gamma)$. It is worth noting that $g(w; \gamma^*) = g^*(w)$. All these considerations lead to the following form of (23)

$$\varepsilon_n(\theta) = H^{-1}(z; \lambda) \Big[Y_n - \Lambda(z; \lambda) g(\gamma^T \mathbf{U}_n; \gamma) \Big]. \tag{24}$$

Reasoning now as in Sect. 2 we can readily form the score function for estimating θ^*

$$Q_N(\theta) = N^{-1} \sum_{n=1}^{N} \varepsilon_n^2(\theta).$$

This is a direct counterpart of the criterion defined in (9). As we have already noted the regression $g(w; \gamma)$ is unknown but can be directly estimated by nonparametric regression estimates. For example, we use the kernel method, see (10),

$$\hat{g}(w; \gamma) = \sum_{t=1}^{N} Y_t K((w - \gamma^T \mathbf{U}_t)/b) / \sum_{l=1}^{N} K((w - \gamma^T \mathbf{U}_l)/b). \tag{25}$$

Using this or any other nonparametric regression estimate in $Q_N(\theta)$ we can form the practical score function for estimating θ^*

$$\hat{Q}_N(\theta) = N^{-1} \sum_{n=1}^{N} \hat{\varepsilon}_n^2(\theta), \tag{26}$$

where $\hat{\varepsilon}_n(\theta)$ is the version of (24) with $g(\gamma^T \mathbf{U}_n; \gamma)$ replaced by $\hat{g}(\gamma^T \mathbf{U}_n; \gamma)$. The minimizer of $\hat{Q}_N(\theta) = \hat{Q}_N(\lambda, \gamma)$ defines an estimate $(\hat{\lambda}, \hat{\gamma})$ of (λ^*, γ^*). Following the reasoning from Sect. 2 we can show that $(\hat{\lambda}, \hat{\gamma})$ tends (P) to (λ^*, γ^*). Furthermore, under additional mild conditions we can find that $(\hat{\lambda}, \hat{\gamma})$ is converging with the optimal $O_P(N^{-1/2})$ rate.

Let us assume that the linear subsystem of the Hammerstein structure is of the finite impulse response type of order p (FIR(p)), i.e., $\Lambda(z; \lambda^*) = 1 + \sum_{i=1}^{p} \lambda_i^* z^{-i}$ [9, 21]. Then we can estimate $\Lambda(z; \lambda^*)$ (independently of $g^*(\bullet)$ and γ^*) via the correlation method, see [13]. In fact, for a given function $\eta : R^d \to R$ such that $E\eta(\mathbf{U}_n) = 0$ and $E\{\eta(\mathbf{U}_n) g^*(W_n)\} \neq 0$ we have the following estimate of λ^*

$$\tilde{\lambda}_t = \frac{N^{-1} \sum_{i=1}^{N-t} Y_{t+i} \eta(\mathbf{U}_i)}{N^{-1} \sum_{i=1}^{N} Y_i \eta(\mathbf{U}_i)}, \quad t = 1, \ldots, p.$$

This applied in (26) gives the simplified least squares criterion for selecting γ

$$\hat{Q}_N(\gamma) = N^{-1} \sum_{i=p+1}^{N} \left(Y_i - \sum_{t=0}^{p} \tilde{\lambda}_t \hat{g}\left(\gamma^T \mathbf{U}_{i-t}; \gamma\right)\right)^2. \tag{27}$$

Once the parametric part of the Hammerstein system is obtained one can define the following nonparametric estimate for the system nonlinearity

$$\hat{g}(w) = \hat{g}(w; \hat{\gamma}),$$

where $\hat{g}(w; \gamma)$ is any nonparametric consistent estimate of $g(w; \gamma)$ and $\hat{\gamma}$ is the minimizer of $\hat{Q}_N(\lambda, \gamma)$.

Recalling the arguments given in Sect. 2 we can conclude that if $g^*(w)$ is twice differentiable and if we select the bandwidth as $b = cN^{-1/5}$ then we have

$$\hat{g}(w) = g^*(w) + O_P(N^{-2/5}).$$

This rate is independent of the dimensionality of the input signal and it is known to be optimal [30]. This should be contrasted with the nonparametric Hammerstein system identification accuracy, see (20). The bandwidth choice is critical for the precision of the kernel estimate. The choice $b = cN^{-1/5}$ is only asymptotically optimal and in practice one would like to specify b depending on the data at hand. One possibility as we have already pointed out, would be to extend the criterion $\hat{Q}_N(\theta)$ in (26) and include b into the minimization process. Hence, we would have the modified criterion $\hat{Q}_N(\theta, b)$.

It is worth noting that in the above two-step scheme the estimate $\hat{g}(w; \gamma)$ in (26) and the criterion function $\hat{Q}_N(\theta)$ share the same training data. This is usually not the recommended strategy since it may lead to estimates with unacceptably large variance. Indeed, some resampling schemes would be useful here which would partition the training data into the testing and training sequences. The former should be used to form the criterion $\hat{Q}_N(\lambda, \gamma)$, whereas the latter to obtain the nonparametric estimate $\hat{g}(w; \gamma)$. The aforementioned concepts are illustrated in the following simulation example, see also [24] for further details.

Example 2. In our simulation example, the d-dimensional input signal $\mathbf{U}_n$ is generated according to uncorrelated Gaussian distribution $N_d(\mathbf{0}, \sigma^2\mathbf{I})$. We assume that the actual system can be exactly represented by the semiparametric model, with the characteristics $\gamma = \left(\cos(\theta), \sin(\theta)/\sqrt{d-1}, \ldots, \sin(\theta)/\sqrt{d-1}\right)^T$ and $g(w) = 0.7 \arctan(\beta w)$. Note that with this parameterization $||\gamma|| = 1$. The true value of γ corresponds to $\theta^* = \pi/4$. The slope parameter β defining $g(w)$ is changed in some experiments. Note that the large β defines the nonlinearity with a very rapid change at $w = 0$. The FIR(3) linear subsystem is used with the transfer function $\Lambda^*(z) = 1 + 0.8z^{-1} - 0.6z^{-2} + 0.4z^{-3}$. The noise Z_t is $N(0, 0.1)$. In our simulation examples we generate L different independent training sets and determine our estimates $\hat{\gamma}$ and $\hat{g}(\cdot)$ described in this section. The local linear kernel estimate with the kernel function $K(w) = (1-w^2)^2$, $|w| \leq 1$ was employed. In implementing the kernel estimate, the window length b was selected simultaneously with the choice of γ. Furthermore, in the partition resampling strategy the size of the training subset is set to 55 % of the complete training data of the size $n = 150$. It is also worth noting that the optimal b needed for estimating a preliminary regression estimate $\hat{g}(w; \gamma)$, has been observed to be different than that required for the final estimate $\hat{g}(w)$.

Figure 5a shows the mean squared error (MSE) of $\hat{\gamma}$ versus the parameter β. Figure 5b gives the identical dependence for the mean integrated squared error (MISE) of $\hat{g}(\cdot)$. In both figures we have $d = 2$. We observe a little influence of the complexity of the nonlinearity $g(w)$ on the accuracy of the estimating γ. This is not the case for estimating $g(w)$. Clearly, a faster changing function is harder to estimate than the one that changes slowly. Figure 6a, b show the influence of the input dimensionality on the accuracy of $\hat{\gamma}$ and $\hat{g}(\cdot)$. The slope parameter is set to $\beta = 2$. As d varies from $d = 2$ to $d = 10$ we observe a very little change in the error values. This supports the observation that the semiparametric approach may behave favorably in high dimensions.

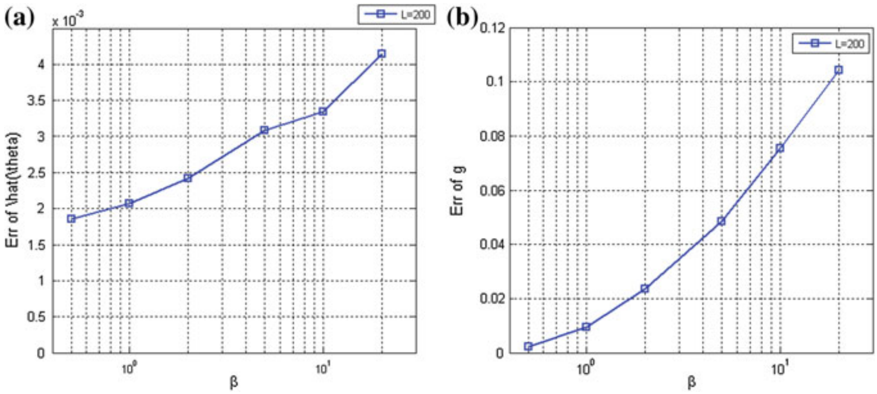


Fig. 5 **a** $MSE(\hat{\gamma})$ versus the slope parameter β; **b** $MISE(\hat{g})$ versus the slope parameter β; $n = 150$, $d = 2$

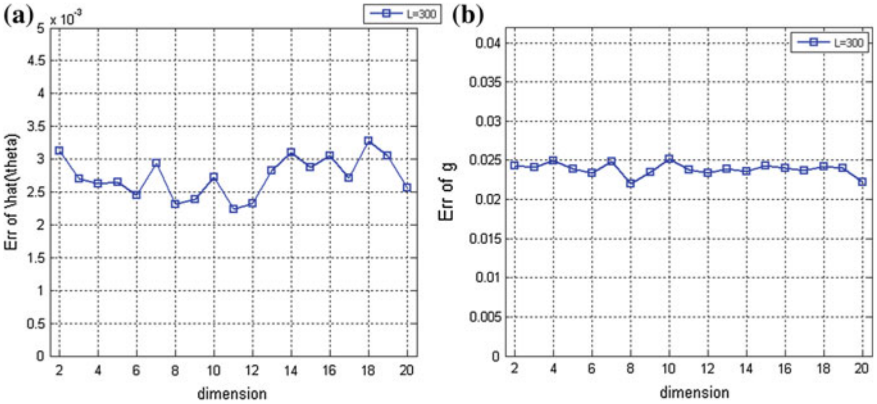


Fig. 6 **a** $MSE(\hat{\gamma})$ versus the input dimensionality d; **b** $MISE(\hat{g})$ versus the input dimensionality d; $n = 150$, $\beta = 2$

3.2 Semiparameric Wiener Systems

In this section, we will illustrate the semiparametric methodology developed in Sect. 2 by the examination of the constrained Wiener system. The system is shown in Fig. 7 and is characterized by the pair $(\lambda^*, m^*(\bullet))$, where $\lambda^* \in R^{p+1}$ is the vector

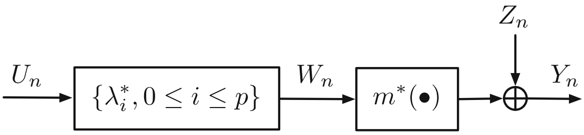


Fig. 7 Semiparametric Wiener system

representing the true impulse response function of the linear subsystem, whereas $m^*(\bullet)$ is a function representing the nonlinear characteristic. The Wiener system can be viewed as the inverse cascade connection to the Hammerstein system examined in the previous section, see [13, 25] for algorithms for nonparametric identification of Wiener systems. Thus, we have the following time-domain input–output relationship:

$$\begin{cases} W_n = \sum_{l=0}^{p} \lambda_l^* U_{n-l} \\ Y_n = m^*(W_n) + Z_n \end{cases}, \tag{28}$$

where the order p of the dynamic FIR(p) type subsystem is assumed to be known. The tandem nature of the Wiener system yields the scaling effect, that is, one can only estimate λ^* up to a multiplicative constant. Let us consider a Wiener system with the characteristics $\bar{m}^*(w) = m^*(w/c)$ and $\bar{\lambda}^* = c\lambda^*$, c being an arbitrary nonzero constant. Then it is easy to see that the new system is indistinguishable from the original one. Thus, in order to get around this identifiability problem we need some normalization of the sequence $\lambda^* = \{\lambda_l^*, 0 \leq l \leq p\}$. A simple normalization is to assume that $\lambda_0 = 1$.

To proceed further, it is necessary to introduce the space Λ of all admissible impulse response functions $\lambda = \{\lambda_l, 0 \leq l \leq p\}$ of order p which satisfy the normalization constraint $\lambda_0 = 1$. Hence, let $\Lambda = \{\lambda \in R^{p+1} : \lambda_0 = 1\}$ such that $\lambda^* \in \Lambda$. By virtue of the semiparametric methodology of Sect. 2, we first wish to characterize the system nonlinearity for a given $\lambda \in \Lambda$. Hence, let

$$W_n(\lambda) = \sum_{l=0}^{p} \lambda_l U_{n-l}, \tag{29}$$

be the interconnecting signal of the Wiener system corresponding to $\lambda \in \Lambda$. Consequently, the following regression function

$$m(w; \lambda) = E\{Y_n | W_n(\lambda) = w\}, \tag{30}$$

plays the role of the best approximate of $m^*(w)$ for a given $\lambda \in \Lambda$. It is clear that $W_n(\lambda^*) = W_n$ and $m(w; \lambda^*) = m^*(w)$. The smoothness of $m(w; \lambda)$ plays an important role in the statistical analysis of our identification algorithms. Since $m(w; \lambda) = E\{m^*(W_n)|W_n(\lambda) = w\}$, the smoothness of $m(w; \lambda)$ is controlled by the smoothness of $m^*(w)$ and the conditional distribution of W_n on $W_n(\lambda)$.

Example 1 To illustrate the dependence of $m(w; \lambda)$ on $m^*(w)$ in terms of smoothness, let $\{U_n\}$ be an *i.i.d* sequence with a normal distribution $N(0, \sigma^2)$. Then, denoting by $\phi(\bullet)$ the $N(0, 1)$ density and after some algebra we have,

$$m(w; \lambda) = \int_{-\infty}^{\infty} m^*(\mu(\lambda)w + v\sigma(\lambda))\phi(v)dv, \tag{31}$$

where

$$\mu(\lambda) = \frac{\lambda^T \lambda^*}{\|\lambda\|^2}, \sigma^2(\lambda) = \sigma^2 \left[\|\lambda^*\|^2 - \frac{(\lambda^T \lambda^*)^2}{\|\lambda\|^2} \right].$$

In Fig. 8 we plot $m(w; \lambda)$ in (31) as a function of w with $\lambda = (1, \lambda_1)^T$, $\lambda^* = (1, 1)^T$, $\sigma^2 = 1$, and the discontinuous nonlinearity $m^*(w) = sgn(w)$. Values $\lambda_1 = -0.9, 0, 0.9$ are used. Note that the value $\lambda_1 = 0$ indicates that there is no dynamical subsystem in the Wiener system. The continuity of $m(w; \lambda)$ is apparent. In Fig. 9, we plot $m(w; \lambda)$ versus λ for a few selected values of w, that is, $w = -1, -0.1, 0.1, 1$. The sensitivity of $m(w; \lambda)$ with respect to λ is small for points which lie far from the point of discontinuity $w = 0$. On the other hand, we observe a great influence of λ at the points which are close to the discontinuity.

In general, we have that

$$m(w; \lambda) = \int_{-\infty}^{\infty} m^*(z) f(z|w; \lambda) dz,$$

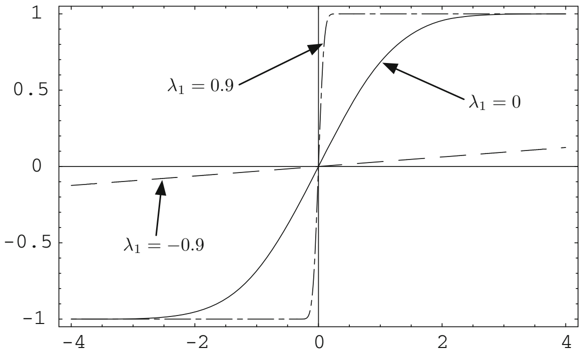


Fig. 8 The regression function $m(w; \lambda)$ in (31) versus w, with $\lambda = (1, \lambda_1)^T$, $\lambda^* = (1, 1)^T$, $m^*(w) = sgn(w)$. Values $\lambda_1 = -0.9, 0, 0.9$

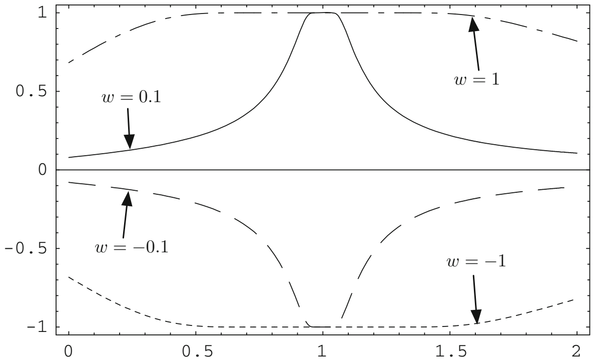


Fig. 9 The regression function $m(w; \lambda)$ in (31) versus λ_1, with $\lambda = (1, \lambda_1)^T$, $\lambda^* = (1, 1)^T$, $m^*(w) = sgn(w)$. Values $w = -1, -0.1, 0.1, 1$

where $f(z|w;\lambda)$ is the conditional probability density function of W_n on $W_n(\lambda)$. Then recalling the result in [22] (Proposition 6.1.1) we can infer that $m(w;\lambda)$ is continuous if

$$w \mapsto \int_A f(z|w;\lambda)dz \tag{32}$$

is lower semicontinuous for every measurable set A. We say that the function $h : R \to R$ is lower semicontinuous if $\liminf_{z\to x} h(z) \geq h(x)$, $x \in R$. Note that the class of lower semicontinuous functions admits some discontinues functions.

Our principal goal is to recover the pair $(\lambda^*, m^*(\bullet))$ from the the training set

$$D_N = \{(U_1, Y_1), \ldots, (U_N, Y_N)\}, \tag{33}$$

where $\{U_i\}$ is a stationary sequence of random variables with the marginal density function $f_U(\bullet)$. Owing to the semiparametric methodology discussed in Sect. 2, we first must estimate the regression function $m(w;\lambda)$. This can easily be done using any previously studied regression estimates applied to synthetic data parametrized by λ:

$$\{(W_{p+1}(\lambda), Y_{p+1}), \ldots, (W_N(\lambda), Y_N)\}. \tag{34}$$

This yields an estimate $\hat{m}(w;\lambda)$, which allows one to define a predictive error as a function of only the linear subsystem characteristic λ. Then, we can write the score function as follows:

$$\hat{Q}_N(\lambda) = N^{-1} \sum_{j=p+1}^{N} (Y_j - \hat{m}(W_j(\lambda);\lambda))^2. \tag{35}$$

The strategy of estimating λ^* is now based on the minimization of $\hat{Q}_N(\lambda)$.

Identification Algorithms The criterion $\hat{Q}_n(\lambda)$ in (35) uses the same data to form the pilot estimate $\hat{m}(w;\lambda)$ and to define $\hat{Q}_n(\lambda)$. This is not generally a good strategy and some form of resampling scheme should be applied in order to separate the data into the testing and training sequences. Hence, consider the partition strategy that reorganizes a set of training data D_N into two non-overlapping subsets that are statistically independent. Owing to the fact that the observations Y_n and Y_{n+p+1} are statistically independent, we define T_1 as the subset of training set D_N consisting of n_1 observations after deleting the first p data points due to the memory effect. Similarly let T_2 be the remaining part of D_N separated from T_1 by the distance of length p. By construction we note that T_1 and T_2 are independent random subsets of D_N. This is the useful property that allows us to design efficient estimates of λ^*, $m(w;\lambda)$, and consequently $m^*(\bullet)$. We use the subset T_1 to estimate the regression function $m(w;\lambda)$ whereas T_2 is used as a testing sequence to form the least-squares

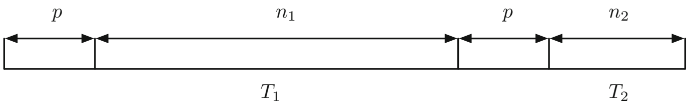


Fig. 10 The partition of the training set D_N into independent subsets T_1 and T_2

criterion to recover the impulse response sequence λ^*. Also, let I_1 and I_2 denote the indices of data points $\{(U_i, Y_i), 1 \le i \le N\}$ which belong to T_1 and T_2, respectively. Figure 10 shows an example of the partition of T into T_1 and T_2. Here we have $n_2 = N - 2p - n_1$.

It is clear that there other possible partitions of a training data set. In fact, the machine learning theory principle says the testing sequence T_2 should consists of independent observations, whereas the training sequence T_1 can be arbitrary [5, 6]. In our situation this strategy can be easily realized by choosing the testing observations that are $p + 1$ positions apart from each other.

The Nadaraya–Watson regression estimate applied to the subset T_1 takes the following form:

$$\hat{m}(w; \lambda) = \frac{\sum_{j \in I_1} Y_j K\left(\frac{w - W_j(\lambda)}{b}\right)}{\sum_{j \in I_1} K\left(\frac{w - W_j(\lambda)}{b}\right)}, \tag{36}$$

for a given $\lambda \in \Lambda$.

This pilot estimate of $m(w; \lambda)$ can now be used to form the least-squares approach to recover the impulse response λ^* of the Wiener system. Thus, the least-squares version of the criterion function in (35) confined to the data set T_2 takes the following form

$$\hat{Q}_N(\lambda) = \frac{1}{n_2} \sum_{i \in I_2} \{Y_i - \hat{m}(W_i(\lambda); \lambda)\}^2. \tag{37}$$

A natural estimate of λ^* is the following minimizer of $\hat{Q}_N(\lambda)$:

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \hat{Q}_N(\lambda). \tag{38}$$

Once the estimate $\hat{\lambda}$ is obtained, one can define the following estimate of the nonlinear characteristic $m^*(\bullet)$ of the Wiener system

$$\hat{m}(w) = \hat{m}(w; \hat{\lambda}). \tag{39}$$

It is clear that the criterion $\hat{Q}_N(\lambda)$ need not possess a unique minimum and, moreover, an efficient procedure to find the minimum of $\hat{Q}_N(\lambda)$ is required. This can be partially overcome by applying the iterative algorithm introduced Sect. 2, see (13). This requires determining for a given $\hat{\lambda}^{(\text{old})}$ the minimum of the partially specified score function,

$$\widetilde{Q}_N(\lambda) = \frac{1}{n_2} \sum_{i \in I_2} \left\{ Y_i - \hat{m}\left(W_i(\lambda); \hat{\lambda}^{(\text{old})} \right) \right\}^2. \tag{40}$$

There are various ways to refine the above procedure by, e.g., employing some preliminary estimate of λ^*, rather than selecting an arbitrary $\hat{\lambda}^{(\text{old})}$ at the initial step. In fact, assuming that the input signal $\{U_n\}$ is a stationary zero-mean white Gaussian process and then by a direct algebra we can show that

$$E\{Y_n U_{n-r}\} = \alpha \lambda_r^*,$$

for $r = 1, 2, \ldots, p$ and where the constant α is well defined assuming that $E|W_0 m^*(W_0)| < \infty$. Since $\lambda_0^* = 1$ we can recover λ_r^* by the following correlation estimate

$$\hat{\lambda}_r = \frac{N^{-1} \sum_{j=r+1}^{N} U_{j-r} Y_j}{N^{-1} \sum_{j=1}^{N} U_j Y_j}, \quad r = 1, \ldots, p.$$

It is worth noting that this is a consistent estimate of λ_r^* provided that the input signal is a stationary white Gaussian process. If the input process is at least close to being Gaussian, we can still choose $\hat{\lambda}_r$ as the initial value $\hat{\lambda}_r^{(\text{old})}$ in the aforementioned iterative algorithm. This may drastically reduce the number of iterations required in the algorithm.

The algorithm uses a kernel estimate which, in turn, needs the selection of the bandwidth parameter b. Due to the splitting strategy, our criterion $\hat{Q}_N(\lambda)$ or its simplified form $\widetilde{Q}_N(\lambda)$ are already in the form of a predictive error, and we can incorporate the bandwidth into the definition of our criterion. Hence, we can use

$$\widetilde{Q}_N(\lambda; b) = \frac{1}{n_2} \sum_{i \in I_2} \left\{ Y_i - \hat{m}\left(W_i(\lambda); \hat{\lambda}^{(\text{old})} \right) \right\}^2$$

as the criterion for selecting both λ and b.

Convergence Analysis This section is concerned with the convergence analysis of the identification algorithms $\hat{\lambda}$ and $\hat{m}(\bullet)$ proposed in (38) and (39), respectively. We will employ the basic methodology established in Sect. 2.

Let $f_W(\bullet)$ and $f_W(\bullet; \lambda)$ be marginal density functions of random processes $\{W_n\}$ and $\{W_n(\lambda)\}$, respectively. Note that $f_W(\bullet)$ and $f_W(\bullet; \lambda)$ always exist since they are obtained by the $(p + 1)$-fold convolution of the scaled version of $f_U(\bullet)$—the probability density function of the input process. In the subsequent sections of this chapter we give sufficient conditions for the convergence of $\hat{\lambda}$ and $\hat{m}(\bullet)$.

A. Parametric Estimation

Owing to the results in Sect. 2, we can extend a definition of the least-squares estimate to a class of minimizers that nearly minimize $\hat{Q}_N(\lambda)$, i.e.,

$$\hat{Q}_N(\hat{\lambda}) \leq \inf_{\lambda \in \Lambda} \hat{Q}_N(\lambda) + \varepsilon_N, \tag{41}$$

for any random sequence $\{\varepsilon_N\}$ such that $\varepsilon_N \overset{N}{\to} 0, (P)$. As we have already noted, (41) implies that $\hat{Q}_N(\hat{\lambda}) \leq \hat{Q}_N(\lambda^*) + \varepsilon_N$ and this is sufficient for the convergence of $\hat{\lambda}$ in (41) to λ^*.

Let us begin with the observation that due to the independence of the sample sets T_1 and T_2 we have,

$$\bar{Q}(\lambda) = E\{\hat{Q}_N(\lambda)|T_1\} = E\{(Y_t - \hat{m}(W_t(\lambda);\lambda))^2|T_1\}, \tag{42}$$

where $(W_t(\lambda), Y_t)$ is a random vector, which is independent of T_1. The definition of $m(w;\lambda)$ in (30) and the fact that the noise is independent of $\{Y_n\}$ yield:

$$\begin{aligned}\bar{Q}(\lambda) = EZ_t^2 &+ E\{(m(W_t(\lambda);\lambda) - m^*(W_t))^2\} \\ &+ E\{(\hat{m}(W_t(\lambda);\lambda) - m(W_t(\lambda);\lambda))^2|T_1\}.\end{aligned} \tag{43}$$

The last term in the above decomposition represents the integrated squared error between the kernel estimate $\hat{m}(w;\lambda)$ and the regression function $m(w;\lambda)$. We can easily show that under the standard assumptions on the kernel function and the bandwidth sequence $\{b_N\}$ (see Assumptions **A4** and **A5** listed below), the last term in (43) tends (P) to zero. Since, moreover, $\hat{Q}_N(\lambda)$ converges (P) to its average $\bar{Q}(\lambda)$ for every $\lambda \in \Lambda$, then we may conclude that:

$$\hat{Q}_N(\lambda) \overset{N}{\to} Q(\lambda), (P) \quad \text{for every } \lambda \in \Lambda, \tag{44}$$

where,

$$Q(\lambda) = EZ_t^2 + E\{(m(W_t(\lambda);\lambda) - m^*(W_t))^2\}. \tag{45}$$

This asymptotic criterion can be now used to characterize the true impulse response function λ^*. In fact, since $Q(\lambda^*) = EZ_t^2$, we have $Q(\lambda^*) \leq Q(\lambda)$, $\lambda \in \Lambda$. Nevertheless, λ^* need not be a unique minimum of $Q(\lambda)$. Indeed, the second term in (45) is equal to zero for such λ values which belong to the following set:

$$S = \{\lambda \in \Lambda : P\{m^*(W_t) = E(m^*(W_t)|W_t(\lambda))\} = 1\}. \tag{46}$$

This set defines all possible values minimizing $Q(\lambda)$ and it is clear that $\lambda^* \in S$. The property $P\{m^*(W_t) = E(m^*(W_t)|W_t(\lambda))\} = 1$ may hold for other λ values, but this happens in very rare cases. Note, however, that $S = R^{p+1}$ if $m^*(\bullet)$ is a constant function. Excluding this singular situation we may certainly assume that $Q(\lambda)$ has the unique global minimum at λ^*. This assumption will be applied throughout our convergence considerations.

The following formal assumptions are required for consistency:

A1 Let the density $f_U(\bullet)$ of the input process be a continuous function bounded away from zero on some small neighborhood of the point $u = 0$.

A2 Let $m^*(\bullet)$ be a non-constant continuous function defined on the support of the random process $\{W_n\}$ such that $E|m^*(W_n)|^2 < \infty$.

A3 Let the space $\Lambda = \{\lambda \in R^{p+1} : \lambda_0 = 1\}$ of all admissible impulse response functions be a compact subset of R^{p+1}.

A4 Let the kernel function $K(\bullet)$ be continuous and satisfy the following restriction:

$$k_1 \mathbf{1}_{[-r,r]}(w) \le K(w) \le k_2 \mathbf{1}_{[-R,R]}(w),$$

for some positive constants $r \le R$, $k_1 \le k_2$.

A5 Let the smoothing sequence $\{b_N\}$ be such that $b_N \to 0$ and $Nb_N \to 0$ as $N \to \infty$.

The kernel function satisfying Assumption **A4** is called a boxed kernel and there is a large class of kernels that may be categorized as such.

The following theorem gives sufficient conditions for the convergence of the identification algorithm defined in (41) to the true impulse response function λ^*.

Theorem 1 *Let $\hat{\lambda}$ be any estimate defined in (41) and let λ^* be a unique minimizer of the limit criterion $Q(\lambda)$. Suppose that Assumptions **A1–A5** hold. Then we have,*

$$\hat{\lambda} \xrightarrow{N} \lambda^*, (P).$$

The critical part in proving this theorem is to show the uniform convergence of $\hat{Q}_N(\lambda)$ to its average $\bar{Q}(\lambda)$, that is, that,

$$\sup_{\lambda \in \Lambda} |\hat{Q}_N(\lambda) - \bar{Q}(\lambda)| \to 0, (P) \quad \text{as } N \to \infty.$$

Such a property is often called a Glivienko–Cantelli property. This is the property of a set of functions,

$$\{(Y - \hat{m}(W(\lambda); \lambda))^2 : \lambda \in \Lambda\}, \tag{47}$$

which defines the criterion $\hat{Q}_N(\lambda)$.

If stronger requirements are imposed on (47), for example, that the nonlinearity $m^*(\bullet)$ and the noise process $\{Z_n\}$ are bounded, then the set in (47) defines the Vapnik–Chervonenkis class. This allows one to show the following exponential inequality:

$$P\left\{\sup_{\lambda \in \Lambda} |\hat{Q}_N(\lambda) - \bar{Q}(\lambda)| \ge \delta | T_1\right\} \le c(N)e^{-\alpha n_2 \delta^2}, \tag{48}$$

for every $\delta > 0$ and some $\alpha > 0$. The sequence $c(N)$ is known to not grow faster than a polynomial in N. It is worth noting that bound (48) holds uniformly over all training sequences T_1 of size n_1. The important consequence of this is that the accuracy of the estimate $\hat{\lambda}$ does not depend critically on the training sequence T_1. Hence, the training sequence can be quite arbitrary, whereas the testing part T_2 of the training set should be as independent as possible [5, 6].

The result of Theorem 1 combined with the further technical arguments mentioned briefly in Sect. 2 allow us to evaluate the rate of convergence of the estimate $\hat{\lambda}$. This is summarized in the following theorem.

Theorem 2 *Let all the assumptions of Theorem 1 be satisfied. Let the derivative $K^{(1)}(\bullet)$ of the kernel function exist and be bounded. Suppose that $f_U(\bullet)$ and $m^*(\bullet)$ have two continuous, bounded derivatives. Then for any sequence of estimators $\hat{\lambda}$ that satisfy (41) with $n\varepsilon_N \overset{N}{\to} 0(P)$ and such that $\hat{\lambda} \overset{N}{\to} \lambda^*(P)$ we have*

$$\hat{\lambda} = \lambda^* + O_P(N^{-1/2}).$$

This result shows that the semiparametric least-squares estimation method can reach the usual $\sqrt{N}$ parametric rate of convergence. Nevertheless, additional smoothness conditions on the input density and system nonlinearity are required. On the contrary, the correlation type estimators of λ^* can reach the $\sqrt{N}$ rate without virtually any assumptions on the nonlinearity and the system memory. The critical assumption, however, is that the input signal is Gaussian [13]. See, however, [12, 25] for alternative nonparametric identification algorithms of Wiener systems without the assumption of Gaussianity.

B. Nonparametric Estimation

The estimate $\hat{\lambda}$ of the linear subsystem found in the preceding section allows one to define an estimate of $\hat{m}(\bullet)$ as in (39), that is, $\hat{m}(\bullet) = \hat{m}(\bullet; \hat{\lambda})$, where $\hat{m}(\bullet; \lambda)$ is the kernel estimate defined in (36). The first step in proving the consistency result for $\hat{m}(\bullet)$ is to apply the decomposition in (16). The convergence of the second term in this decomposition

$$\hat{m}(\bullet; \lambda^*) - m^*(\bullet) \overset{N}{\to} 0, (P), \tag{49}$$

represents the classical problem in nonparametric estimation. In our case the output process is p-dependent, that is, the random variables Y_i and Y_j are independent as long as $|i - j| > p$. Then the proof of (49) results from the fact that for any p-dependent random process $\{\xi_i\}$ such that $E\{\xi_i\} = 0$ and $E\xi_i^2 < \infty$ we have

$$E\left(\sum_{j=1}^{N} \xi_j\right)^2 \leq (p+1) \sum_{j=1}^{N} E\xi_j^2. \tag{50}$$

Concerning the first term in (16), note that we wish to apply the linearization technique with respect to $\hat{\lambda} - \lambda^*$. To do so, let us write the kernel estimate $\hat{m}(w; \lambda)$ in (36) as follows:

$$\hat{m}(w; \lambda) = \frac{\hat{r}(w; \lambda)}{\hat{f}(w; \lambda)}, \tag{51}$$

where

$$\hat{r}(w; \lambda) = n_1^{-1} b^{-1} \sum_{j \in I_1} Y_j K\left(\frac{w - W_j(\lambda)}{b}\right)$$

and

$$\hat{f}(w; \lambda) = n_1^{-1} b^{-1} \sum_{j \in I_1} K\left(\frac{w - W_j(\lambda)}{b}\right).$$

Note that $\hat{f}(w; \lambda)$ is the kernel estimate of the density function $f(w; \lambda)$, whereas $\hat{m}(w; \lambda)$ is the kernel estimate of $m(w; \lambda) f(w; \lambda)$.

Now using (51) and recalling that $W_j(\lambda^*) = W_j$, we can express the derivative of $\hat{m}(w; \lambda^*)$ with respect to $W_j,\ j \in I_1$ as follows:

$$D_j(w) = n_1^{-1} b^{-2} K^{(1)}\left(\frac{w - W_j}{b}\right) \cdot \frac{\hat{r}(w; \lambda^*) - Y_j \hat{f}(w; \lambda^*)}{\hat{f}^2(w; \lambda^*)}, \tag{52}$$

where $\hat{r}(w; \lambda^*)$, $\hat{f}(w; \lambda^*)$ are defined as in (51) with $\lambda = \lambda^*$. Next, let us note that

$$W_j(\hat{\lambda}) - W_j(\lambda^*) = \sum_{t=1}^{p} (\hat{\lambda}_t - \lambda_t^*) U_{j-t}, \quad j \in I_1.$$

Then we can approximate $\hat{m}(w) - \hat{m}(w; \lambda^*)$ by the first term of Taylor's formula,

$$\sum_{j \in I_1} D_j(w)(W_j(\hat{\lambda}) - W_j(\lambda^*)) = \sum_{t=1}^{p} (\hat{\lambda}_t - \lambda_t^*) A_{t,n}(w),$$

where

$$A_{t,N}(w) = \sum_{j \in I_1} D_j(w) U_{j-t},$$

for $1 \le t \le p$.

Since, by Theorem 1, we have that $\hat{\lambda}_t - \lambda_t^* \overset{N}{\to} 0(P)$, it is sufficient to show that the stochastic term $A_{t,N}(w)$ tends (P) to a finite function as $N \to \infty$. Let us note that by the standard technical considerations we can show that $\hat{f}(w; \lambda^*)$ and $\hat{r}(w; \lambda^*)$ converge (P) to $f_W(w)$ and $m^*(w) f_W(w)$, respectively. By this and (52), we see that the term $A_{t,N}(w)$ is determined by the following two expressions:

$$J_1(w) = n_1^{-1} b^{-2} \sum_{j \in I_1} K^{(1)} \left(\frac{w - W_j}{b} \right) U_{j-t},$$

$$J_2(w) = n_1^{-1} b^{-2} \sum_{j \in I_1} K^{(1)} \left(\frac{w - W_j}{b} \right) Y_j U_{j-t}.$$

It suffices to examine the term $J_2(w)$. Let us start by noting that

$$J_2(w) = \frac{\partial}{\partial w} \bar{J}_2(w), \tag{53}$$

where

$$\bar{J}_2(w) = n_1^{-1} b^{-1} \sum_{j \in I_1} K \left(\frac{w - W_j}{b} \right) Y_j X_{j-t}.$$

It can be shown by using (50) that

$$\bar{J}_2(w) \overset{N}{\to} m^*(w) a(w), (P), \tag{54}$$

where $a(w)$ is some finite function. The convergence (P) of $\bar{J}_2(w)$ implies the convergence (P) of the derivative due to the general result presented in [35]. Thus, by (53) and (54) we obtain

$$J_2(w) \overset{N}{\to} \frac{\partial}{\partial w} \{m^*(w) a(w)\}, (P).$$

The aforementioned discussion explains the main steps used to prove the convergence of the estimate $\hat{m}(w)$ defined in (39) to the true nonlinearity $m^*(w)$. Note that the linearization technique requires some differentiability conditions both on the system characteristics and the kernel function. Hence, we need the following additional formal assumptions:

A6 Let $f_U(\bullet)$ have a bounded and continuous derivative.
A7 Let $m^*(\bullet)$ have a bounded and continuous derivative.
A8 Let the derivative $K^{(1)}(\bullet)$ of the kernel function exist and be bounded.

All these considerations lead to the following convergence result for the nonlinear subsystem identification algorithm.

Theorem 3 *Let $\hat{m}(\bullet) = \hat{m}(\bullet; \hat{\lambda})$, where $\hat{m}(\bullet; \lambda)$ is the kernel regression estimate defined in (36). Let all of the assumptions of Theorem 1 hold. If, further, Assumptions **A6–A8** are satisfied, then we have*

$$\hat{m}(w) \to m^*(w), (P), \quad \textit{as } N \to \infty$$

at every point $w \in R$ where $f_W(w) > 0$.

The conditions imposed in Theorem 3 are by no means the weakest possible and it may be conjectured that the convergence holds at a point where $f_W(w)$ and $m^*(w)$ are continuous.

In the proof of Theorem 3, we have already shown that $\hat{m}(w) - \hat{m}(w, \lambda^*)$ is of order,

$$\sum_{t=1}^{p} (\hat{\lambda}_t - \lambda_t)\, A_{t,N}(w),$$

where $A_{t,N}(w) \overset{N}{\to} A_t(w)(P)$, some finite function $A_t(w)$. Then, due to Theorem 2, we have that

$$\hat{m}(w) - m^*(w) = \{\hat{m}(w; \lambda^*) - m^*(w)\} + O_P(N^{-1/2}). \tag{55}$$

Hence, the rate of convergence of $\hat{m}(w)$ to $m^*(w)$ depends merely on the speed at which the first term on the right-hand side of (55) tends to zero. This is, however, an usual problem in nonparametric estimation. Indeed, the rate is controlled by the smoothness of the nonlinearity $m^*(w)$ and density $f_W(\bullet)$. Note that the smoothness of $f_W(\bullet)$ can be inferred by the smoothness of $f_U(\bullet)$. Since we have assumed that $f_U(\bullet)$ and $m^*(\bullet)$ have two continuous bounded derivatives, then by standard analysis we may readily obtain that $\hat{m}(w; \lambda^*) - m^*(w) = O_P(N^{-2/5})$, provided that the kernel function is even and the bandwidth b is selected as $b(N) = aN^{-1/5}$, some positive constant a. Consequently, we come to the following theorem.

Theorem 4 *Let all the assumptions of Theorems 2 and 3 be satisfied. Suppose that the kernel function is even. Then we have*

$$\hat{m}(w) = m^*(w) + O_P(N^{-2/5}).$$

Extensions Thus far, we have examined the one channel Wiener system with a finite memory and the univariate nonlinearity. We have employed the semiparametric approach to identify the parametric and nonparametric parts of the system. This strategy can be further extended to other types of Wiener systems. Among many possible alternatives we single out a multichannel model with separate dynamical parts and a common nonlinearity. A two-channel version of this particular class of Wiener systems is shown in Fig. 11. This model is important since the celebrated result due to Boyd and Chua [3] assures that any time-invariant nonlinear system, which satisfies the so-called fading memory property, can be approximated by a nonlinear moving-average operator having the structure depicted in Fig. 11. In the statistical setting the fading memory assumption results in a certain type of physical/predictive data dependence such that two input signals which are close in the recent past, but not necessarily close in the remote past, yield present outputs that are close. A similar concept of dependence has been introduced in [32], see also [33].

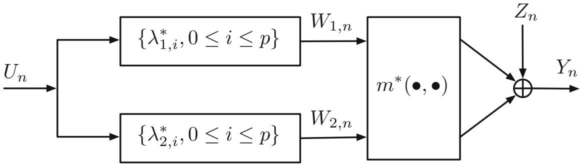

U_n
$\{\lambda^*_{1,i}, 0 \le i \le p\}$
$\{\lambda^*_{2,i}, 0 \le i \le p\}$
$W_{1,n}$
$W_{2,n}$
$m^*(\bullet, \bullet)$
Z_n
Y_n

Fig. 11 The generalized two-channel Wiener model

The model in Fig. 11 can be easily identified within the semiparametric framework examined in Sect. 2. Hence, without loss of generality, let us consider the following two-channel system

$$Y_n = m^* \left(\sum_{i=0}^{p} \lambda^*_{1,i}\, U_{n-i}, \sum_{j=0}^{p} \lambda^*_{2,j}\, U_{n-j} \right) + Z_n, \tag{56}$$

where $\lambda^*_1 = \{\lambda^*_{1,i}, 0 \le i \le p\}$ and $\lambda^*_2 = \{\lambda^*_{2,i}, 0 \le i \le p\}$ are unknown parameters and $m^*(\bullet, \bullet)$ is the unknown nonlinearity.

The first important issue, similar to that studied for the single input Wiener model, is whether the parameter $\lambda^* = (\lambda^*_1, \lambda^*_2) \in R^s$, $s = 2p + 2$, is identifiable. The previous normalization $\lambda_{1,0} = \lambda_{2,0} = 1$ is not sufficient in this case; we must further restrict a class of admissible impulse response sequences and nonlinearities. Concerning the parameter space of all admissible impulse response functions we assume that $\lambda \in \Lambda \subset R^s$ for Λ being a compact subset of R^s, where $\lambda = (\lambda_1, \lambda_2)$.

In general, we can only identify a linear subspace spanned by $(\lambda^*_1, \lambda^*_2)$. To be able to identify the individual parameters we can assume that λ^*_1 and λ^*_2 are not collinear. Furthermore, assume that $m^*(\bullet, \bullet)$ is not a constant function and that the derivatives of $m^*(w_1, w_2)$ with respect to each of the variables are not linearly dependent. This assures us that the nonlinearity is sufficiently far from being constant and linear.

The solution of the identification problem for the model (56) is now straightforward. Indeed, we can follow the ideas developed in the previous section starting with an important concept of the optimal predictor of the output signal for a given $\lambda \in \Lambda$,

$$m(w_1, w_2; \lambda) = E\{Y_n | W_{1,n}(\lambda_1) = w_1, W_{2,n}(\lambda_2) = w_2\}, \tag{57}$$

where $W_{1,n}(\lambda_1) = \sum_{i=0}^{p} \lambda_{1,i} U_{n-i}$ and $W_{2,n}(\lambda_2) = \sum_{j=0}^{p} \lambda_{2,j} U_{n-j}$. We have the obvious constraints $W_{1,n}(\lambda^*_1) = W_{1,n}$, $W_{2,n}(\lambda^*_2) = W_{2,n}$ and $m(\bullet, \bullet; \lambda^*) = m^*(\bullet, \bullet)$. Next, using the partition strategy of the training set shown in Fig. 10, the regression function $m(w_1, w_2; \lambda)$ can be estimated by the two-dimensional version of the kernel estimate $\hat{m}(w_1, w_2; \lambda)$ for a given $\lambda = (\lambda_1, \lambda_2) \in \Lambda$. This allows us to obtain the least-squares score function estimate $\hat{\lambda}$ of λ^* having the $\sqrt{N}$ convergence property. The corresponding estimate of $m^*(w_1, w_2)$ is $\hat{m}(w_1, w_2) = \hat{m}(w_1, w_2; \hat{\lambda})$. Then the reasoning leading to the results of Theorems 3 and 4 readily yields,

$$\hat{m}(w_1, w_2) = m^*(w_1, w_2) + O_P(N^{-1/3}), \tag{58}$$

where $f_U(\bullet)$ and $m^*(\bullet, \bullet)$ are twice continuously differentiable. The rate in (58) needs the proper choice of the bandwidth parameter $b(N)$ of the following form $b(N) = aN^{-1/6}$, for some positive constant a. Note that the rate in (58) is slower that that for the one channel Wiener system, see Theorem 4. This is due to the fact that we are estimating a bivariate function for which the rate is slower than for an univariate one. Further restriction of $m^*(w_1, w_2)$ to the class of additive functions of the form $m_1(w_1) + m_2(w_2)$ would yield the identification algorithm with the one-dimensional rate $O_P(N^{-2/5})$, see [13] for some results regarding identification of additive systems.

3.3 Semiparametric Parallel Systems

In this section we make use of the semiparametric methodology in the context of the parallel system with a single (without loss of generality) input and a finite memory linear subsystem. Hence, the system shown in Fig. 12 is assumed to be the true system with the following input-output description:

$$Y_n = m^*(U_n) + \sum_{j=0}^{p} \lambda_j^* U_{n-j} + Z_n. \tag{59}$$

The identifiability condition for this system is that $\lambda_0^* = 1$. Hence, let $\Lambda = \{\lambda \in R^{p+1} : \lambda_0 = 1\}$ be a set of all admissible parameters that is assumed to be the compact subset of R^{p+1}.

As we have already discussed, the semiparametric least squares strategy begins with the elimination of the nonlinear characteristic from the optimization process. To this end let,

$$W_n(\lambda) = \sum_{j=0}^{p} \lambda_j U_{n-j}, \tag{60}$$

be the output of the linear subsystem for a given $\lambda \in \Lambda$. Clearly $W_n(\lambda^*) = W_n$.

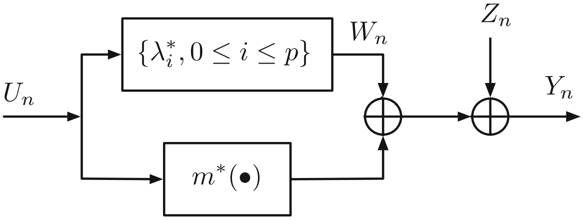


Fig. 12 Semiparametric nonlinear parallel model

Next, we form the required projection

$$m(u;\lambda) = E\{Y_n - W_n(\lambda)|U_n = u\} \tag{61}$$

such that $m(u;\lambda^*) = m^*(u)$. In fact, noting that

$$m(u;\lambda) = m^*(u) + \sum_{j=0}^{p}(\lambda_j^* - \lambda_j)\ E\{U_{n-j}|U_n = u\},$$

we can confirm that $m(u;\lambda^*) = m^*(u)$.

For a given training set $D_N = \{(U_1, Y_1), \ldots, (U_N, Y_N)\}$ we can easily form a nonparametric estimate of the regression function $m(u;\lambda)$. Hence let,

$$\hat{m}(u;\lambda) = \frac{\sum_{t=p+1}^{N}(Y_t - W_t(\lambda))K\left(\frac{u-U_t}{b}\right)}{\sum_{t=1}^{N} K\left(\frac{u-U_t}{b}\right)}, \tag{62}$$

be the kernel regression estimate of $m(u;\lambda)$.

The mean-squared criterion for estimating λ^* can now be defined as follows:

$$\hat{Q}_N(\lambda) = N^{-1}\sum_{t=p+1}^{N}\left(Y_t - \hat{m}(U_t;\lambda) - W_t(\lambda)\right)^2. \tag{63}$$

The minimizer of the prediction error $\hat{Q}_N(\lambda)$ defines an estimate $\hat{\lambda}$ of λ^*. As soon as $\hat{\lambda}$ is determined we can estimate $m^*(u)$ by the two-stage process, i.e., we have,

$$\hat{m}(u) = \hat{m}(u;\hat{\lambda}). \tag{64}$$

Thus far we have used the same data for estimating the pilot regression estimate $\hat{m}(u;\lambda)$ and the criterion function $\hat{Q}_n(\lambda)$. This may lead to consistent estimates but the mathematical analysis of such algorithms is lengthy. In Sect. 3.2 we suggested the partition resampling scheme which gives a desirable separation of the training and testing data sets and reduces the mathematical complications. This strategy can be easily applied here, i.e., we can use a subset of D_N to derive the kernel estimate in (62) and then utilize the remaining part of D_N for computing the criterion function $\hat{Q}_N(\lambda)$.

For estimates of $\hat{\lambda}$ and $\hat{m}(u)$ obtained as outlined above, we can follow the arguments given in Sect. 2 and show that $\hat{\lambda} \rightarrow \lambda^*(P)$ and consequently $\hat{m}(u;\hat{\lambda}) \rightarrow m(u;\lambda^*) = m^*(u)(P)$.

The minimization procedure required to obtain $\hat{\lambda}$ can be involved due to the highly nonlinear nature of $\hat{Q}_N(\lambda)$. A reduced complexity algorithm can be developed based on the general iterative scheme described in Sect. 2, see (13). Hence, for a given $\hat{\lambda}^{(\mathrm{old})}$, set $\hat{m}(u;\hat{\lambda}^{(\mathrm{old})})$. Then we form the modified criterion,

$$\widetilde{Q}_N(\lambda) = N^{-1} \sum_{t=p+1}^{N} \left(Y_t - \hat{m}(U_t; \hat{\lambda}^{(\mathrm{old})}) - W_t(\lambda) \right)^2, \tag{65}$$

and find

$$\hat{\lambda}^{(\mathrm{new})} = \arg\min_{\lambda \in \Lambda} \widetilde{Q}_N(\lambda).$$

Next, we use $\hat{\lambda}^{(\mathrm{new})}$ to get $\hat{m}(u; \hat{\lambda}^{(\mathrm{new})})$ and iterate the above process until the criterion $\widetilde{Q}_N(\lambda)$ does not change significantly. It is worth noting that $W_t(\lambda)$ in (65) is a linear function of λ and therefore we can explicitly find $\hat{\lambda}^{(\mathrm{new})}$ that minimizes $\widetilde{Q}_N(\lambda)$.

We should note that the above algorithm can work with the dependent input process $\{U_n\}$. However, if $\{U_n\}$ is a sequence of $i.i.d.$ random variables, then the correlation method provides the following explicit solution for recovering λ^*. In fact, we have

$$\lambda_j^* = \frac{cov(Y_n, U_{n-j})}{var(U_0)}; \qquad j = 1, \ldots, p.$$

Note also that

$$m^*(u) = E\{Y_n | U_n = u\} - u.$$

which allows us to recover $m^*(u)$. Empirical counterparts of $cov(Y_n, U_{n-j})$, $var(U_0)$, and the regression function $E\{Y_n|U_n = u\}$ define the estimates of the system characteristics. Although these are explicit estimates, they are often difficult to generalize in more complex cases. On the other hand, the semiparametric approach can easily be extended to a large class of interconnected complex systems.

4 Concluding Remarks

In this paper we have described the unified framework for identification of systems that can be represented or approximated by the infinite dimensional parameter and a set of univariate nonparametric functions. The parametric part of this paradigm is representing linear dynamic subsystems of the nonlinear model as well as projections on the low-dimensional subspaces. The latter case is essential for reducing the curse of dimensionality. The developed methodology is illustrated in the context of popular block-oriented systems. We have argued that the semiparametric inference can offer an attractive strategy for identification of large scale composite systems where one faces an inherent problem of dimensionality and model complexity. In fact, the semiparametric paradigm allows us to project the original system onto some parsimonious alternative. The semiparametric version of the least squares method employed in this paper determines such a projection via an optimization procedure. We have examined a class of semiparametric dynamic systems characterized by functions of single variable and finite dimensional projection parameters. The model in (3) is an important example of this class. The following is the natural generalization of the approximation in (3)

$$\mu(\mathbf{x}) = \sum_{l=1}^{L} g_l(\theta_l^T \mathbf{x}), \tag{66}$$

where now we wish to specify the univariate functions $\{g_l(\bullet), 1 \leq l \leq L\}$ and the parameters $\{\theta_l, 1 \leq l \leq L\}$. Often one also needs to estimate the degree L of this approximation network. The approximation properties of (66) has been examined in [7]. It is worth noting the nonlinear characteristic in Example 1 of Sect. 2, i.e., $m(x_1, x_1) = x_1 x_2$, can be exactly reproduced by the network in (66). In fact, we have

$$x_1 x_2 = \frac{1}{4}(x_1 + x_2)^2 - \frac{1}{4}(x_1 - x_2)^2.$$

This corresponds to (66) with $g_1(w) = \frac{1}{4}w^2$, $g_2(w) = -\frac{1}{4}w^2$ and $\theta_1 = (1, 1)^T$, $\theta_2 = (1, -1)^T$.

Semiparametric models have been extensively examined in the econometric literature, see [15, 20, 34]. There, they have been introduced as more flexible extension of the standard linear regression model and popular models include partial linear and multiple-index models. These are static models and this paper can be viewed as the generalization of these models to dynamic nonlinear block-oriented systems. In fact, the partially linear models fall into the category of parallel models, whereas multiple-index models correspond to Hammerstein/Wiener connections. Semiparametric models have recently been introduced in the nonlinear time series literature [9, 27]. Some preliminary results on semiparametric inference in system identification are reported in [8, 13].

In the approximation theory the model in (3) has been recently examined, see [4] and the references cited therein, as the one-dimensional approximation to functions of many variables. The problem of learning of such approximations from a finite number of point queries has been studied. The accuracy of such approximations depends on the smoothness of $g(\bullet)$ and the sparsity of the vector $\theta \in R^q$.

There are numerous ways to refine the results of this paper. First of all, one can consider a more robust version the least-square criterion with a general class of loss function. This would lead to the semiparametric alternative of M-estimation [28]. As a result, we could examine semiparametric counterparts of maximum-likelihood estimation and some penalized M-estimators. The latter would allow us to incorporate some shape constraints like convexity and monotonicity of underlying characteristics. The extension of the semiparametric strategy to continuous-time systems would be an interesting problem for future research. The issue of identification of highly dimensional dynamic systems modelled by semiparametric structures with some sparsity constraints could be another area of interest. On the more technical side, the question of finding semiparametric efficient estimators of the parametric component of a semiparametric block-oriented model remains an open issue, see [28] for the basic theory of semiparametric efficiency.

Acknowledgments The author wishes to thank Mount First for assistance. We would also like to thank the referee for his valuable comments and corrections.

References

1. Andrews DWK (1984) Non-strong mixing autoregressive process. J Appl Probab 21:930–934
2. Billings S (2013) Nonlinear system identification. Wiley, New York
3. Boyd S, Chua L (1985) Fading memory and the problem of approximating nonlinear operators with Volterra series. IEEE Trans Circuits Syst 32:1150–1161
4. Cohen A, Daubechies I, DeVore R, Kerkyacharian G, Picard D (2012) Capturing ridge functions in high dimensions from point queries. Contr Approx 35:225–243
5. Devroye L (1988) Automatic pattern recognition: a study of the probability of error. IEEE Trans Pattern Anal Mach Intell 10:530–543
6. Devroye L, Györfi L, Lugosi G (1996) A probabilistic theory of pattern recogntion. Springer, New York
7. Diaconis P, Shahshahani M (1984) On nonlinear functions of linear combinations. SIAM J Sci Comput 5(1):175–191
8. Espinozo M, Suyken JAK, De Moor B (2005) Kernel based partially linear models and nonlinear identification. IEEE Trans Autom Contr 50:1602–1606
9. Fan J, Yao Q (2003) Nonlinear time series: nonparametric and parametric methods. Springer, New York
10. Giannakis GB, Serpendin E (2001) A bibliography on nonlinear system identification. Sig Process 81:533–580
11. Giri F, Bai EW (eds) (2010) Block-oriented nonlinear system identification. Springer, New York
12. Greblicki W (2010) Nonparametric input density-free estimation of nonlinearity in Wiener systems. IEEE Trans Inform Theory 56:3575–3580
13. Greblicki W, Pawlak M (2008) Nonparametric system identification. Cambridge University Press, Cambridge
14. Härdle W, Hall P, Ichimura H (1993) Optimal smoothing in single-index models. Ann Stat 21:157–178
15. Härdle W, Müller M, Sperlich S, Werwatz A (2004) Nonparametric and semiparametric models. Springer, New York
16. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York
17. Isermann R, Munchhof M (2011) Identification of dynamic systems: an introduction with applications. Springer, New York
18. Koronacki J, Ćwik J (2008) Statystyczne systemy uczace sie (in Polish). Exit, Warsaw
19. Kvam PH, Vidakovic B (2007) Nonparametric statistics with applications to science and engineering. Wiley, New York
20. Li Q, Racine JS (2007) Nonparametric econometrics. Princeton University Press, Princeton
21. Ljung L (1999) System identification: theory for the user. Prentice-Hall, Englewood Cliffs
22. Meyn SP, Tweedie RL (1993) Markov chain and stochastic stability. Springer, New York
23. Mohri M, Rostamizadeh A, Talwalker A (2012) Foundations of machine learning. The MIT Press, Cambridge
24. Pawlak M, Lv J (2014) On nonparametric identification of MISO Hammerstein systems
25. Pawlak M, Hasiewicz Z, Wachel P (2007) On nonparametric identification of Wiener systems. IEEE Trans Signal Process 55:482–492
26. Pillonetto G, Dinuzzo F, Che T, De Nicolao G, Ljung L (2014) Kernel methods in system identification, machine learning and function estimation: a survey. Automatica 50:657–682

27. Saart P, Gao J, Kim NH (2014) Semiparametric methods in nonlinear time series: a selective review. J Nonparametric Stat 26:141–169
28. van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge
29. Vidyasagar M, Karandikar RL (2008) A learning theory approach to system identification and stochastic adaptive control. J Process Contr 18:421–430
30. Wasserman L (2006) All of nonparametric statistics. Springer, New York
31. Westwick D, Kearney R (2003) Identification of nonlinear physiological systems. Wiley, New York
32. Wu WB (2005) Nonlinear system theory: another look at dependence. Proc Nat Acad Sci 102:14150–14154
33. Wu WB, Mielniczuk J (2010) A new look at measuring dependence. In: Doukham P et al. (eds) Dependence in probability and statistics. Springer, New York, pp 123–142
34. Yatchev A (2003) Semiparametric regression for the applied econometrician. Cambridge University Press, Cambridge
35. Yatracos Y (1989) On the estimation of the derivative of a function with the derivative of an estimate. J Multivar Anal 28:172–175

Dealing with Data Difficulty Factors While Learning from Imbalanced Data

Jerzy Stefanowski

Abstract Learning from imbalanced data is still one of challenging tasks in machine learning and data mining. We discuss the following data difficulty factors which deteriorate classification performance: decomposition of the minority class into rare sub-concepts, overlapping of classes and distinguishing different types of examples. New experimental studies showing the influence of these factors on classifiers are presented. The paper also includes critical discussions of methods for their identification in real world data. Finally, open research issues are stated.

1 Introduction

Data mining and machine learning have shown tremendous progress in last decades and have become ones of the main sub-fields of computer sciences. The supervised learning of object classification is one of the most common tasks considered both in theory and practice. Discovered classification knowledge is often used as a classifier to predict class labels for unclassified, new instances. This task has been intensively studied and a large number of approaches, based on different principles, have been already introduced; for some reviews the reader can consult, e.g. [4, 49].

Nevertheless many real world problems still reveal difficulties for learning accurate classifiers and require new solutions. One of these challenges is *learning from imbalanced data*, where at least one of the target classes contains a much smaller number of examples than the other classes. This class is usually referred to as the *minority class*, while the remaining classes are denoted as *majority ones*. For instance, in medical problems the number of patients requiring special attention is much smaller than the number of patients who do not need it. Class imbalances have been also observed in many other application domains such as fraud detection in telephone calls or credit cards transactions, bank risk analysis, technical diagnostics, network intrusion detection, image recognition, detecting specific astronomical objects in sky surveys,

J. Stefanowski (✉)
Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
e-mail: jerzy.stefanowski@cs.put.poznan.pl

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_17

text categorization, information filtering; for some reviews see, e.g., [10, 18, 29, 30, 72].

In all those problems, the correct recognition of the minority class is of key importance. For instance, in a medical diagnosis a failure in recognizing an illness and not assigning a proper treatment to a patient is much more dangerous than misdiagnosing a healthy person, whose diagnosis may be verified by additional examinations. Although focusing attention on a critical class and considering misclassification errors are similar to cost sensitive learning [16], dealing with imbalanced classes is not the same, as the costs of misclassifications are rather unknown in advance [50]. Even if they could be somehow approximated, they may be different for particular instances of the given class.

The standard classifiers do not work sufficiently well with imbalanced data [29, 30, 41, 74]. They mainly concentrate on larger classes and often fail to classify sufficiently accurately minority class examples. For instance, [45] describes an information retrieval system, where the minority class contained only 0.2 % of all examples. Although all considered classifiers achieved the overall accuracy close to 100 %, they were useless because they failed to deliver requested documents from this class. While this degradation of classification performance has been known earlier from applications, improving classifiers for imbalanced data has received a growing research interest in the last decade and a number of specialized methods have been proposed, for their review see, e.g., [10, 18, 29, 30].

Although several specialized methods exist, the identification of conditions for their efficient use is still an open research problem. It is also related to more fundamental issues of better understanding the nature of the imbalance data and key properties of its underlying distribution which makes this problem too difficult to be handled.

Note that many authors introducing their new method usually carry out its experimental evaluation over some data sets and show its better performance than some reference methods. However, these evaluations are usually quite limited and authors do not ask the above mentioned questions on data characteristics. In more comprehensive comparative studies, as [2, 70], data sets are categorized with respect to the global ratio between imbalanced classes or the size of the minority class only. Nevertheless, it seems that these numbers do not sufficiently explain differences between classification performance of the compared methods. For instance, for some data sets even with a high imbalance ratio, the minority class can be sufficiently recognized by many standard classifiers.

Some researchers claim that the global imbalance ratio is not a problem itself and it may not be the main source of difficulties for learning classifiers. Following related works [22, 34, 37, 47, 60] and earlier studies of Stefanowski et al. [52, 53, 55, 65] we claim that other, as we call them, *data difficulty factors*, referring to characteristics of minority class distributions, are also influential. They include:

- decomposition of the minority class into many rare sub-concepts—also known as small disjuncts [32, 34, 36, 67, 73],
- an effect of too strong overlapping between the classes,
- a presence of too many minority examples inside the majority class region.

When these data difficulty factors occur *together* with class imbalance, they may seriously hinder the recognition of the minority class, see e.g. a study [47, 64]. Moreover, in earlier paper of Stefanowski et al. we have proposed to capture some of these data difficulty factors by considering the local characteristics of learning examples from the minority class [53, 55].

We claim that the studies on data difficulty factors are still not sufficiently developed and even well known among machine learning or data mining communities. Furthermore, most of these studies have been carried out on special synthetic data with assumed distributions of the minority class, while good identification of these factors in the real data sets is not easy and it poses still open questions and requires new solutions.

The aim of this paper is to survey the main current research on the above mentioned data difficulty factors including our own new experimental results. We will present consequences of these data factors on the classification performance. Then, we critically discuss current methods for their identification and put open questions on the directions of their future developments. Finally, we will claim that the proper analyzing of these data factors could be the basis for developing new specialized algorithms for imbalanced data.

The paper is organized as follows. Section 2 summarizes related pre-processing methods and evaluation issues. Difficulties caused by a fragmentation of the minority class into rare sub-concept are described in Sect. 3. It is followed by a discussion of class overlapping in Sect. 4. Then, the novel view of types of minority examples, the method for their identification in real world data sets, its experimental evaluation are presented. The final section concludes the paper.

2 Pre-processing Methods for Class Imbalances

Methods addressing class imbalances are usually categorized into two groups:

- **Data level methods**—these are classifier-independent methods that rely on transforming the original data distribution of classes into the better one for learning classifiers, e.g., by re-sampling or focused filtering some examples.
- **Algorithmic level methods**—they involve modifications of the algorithm.

In this paper we do not intend to provide a comprehensive review of all proposed methods and rather will briefly present the selected data level methods only as they will be considered in further experiments. The comprehensive reviews can be found in, e.g., [10, 18, 29, 30, 52, 66, 72].

The methods on the algorithmic level include the following sub-categories: adaptations to cost-sensitive learning, changing of internal algorithm bias (either in search strategies, evaluation criteria or classification strategies), generalizations of ensembles or one-class learning. On the other hand, methods on data level modify imbalanced data to provide the class distribution more appropriated for learning classifiers. Many of these proposed methods offer a more balanced distribution of classes.

In general, changing the class distribution towards a more balanced one improves the performance for most data sets and classifiers [29]. We describe the best well known pre-processing methods below.

2.1 Random Re-sampling and Informed Pre-processing

The most popular re-sampling methods are random *over-sampling* which replicates examples from the minority class, and random *under-sampling* which randomly eliminates examples from the majority classes until a required degree of balance between class cardinalities is reached. However, several authors showed the simple random re-sampling methods were not sufficiently good at improving recognition of imbalanced classes. Random under-sampling may potentially remove some important examples and simple over-sampling may also lead to overfitting [11, 42]. The recent research focuses on particular examples, taking into account information about their distribution in the attribute space [29].

Kubat and Matwin claim in [42] that characteristics of mutual positions of examples is a source of difficulty for learning from imbalanced data, see also their more applied study [43]. They introduced *one-side-sampling* method (OSS), which filters the majority classes in a focused way [42]. It is based on distinguishing different types of learning examples: safe examples, borderline (located near the decision boundary) and noisy examples. They propose to use Tomek links (two nearest examples having different labels) to identify and delete the borderline and noisy examples from majority classes.

Then, the *Nearest Cleaning Rule* (NCR) method is introduced in [44] and it is based on the focused removal of examples from the majority class. It applies the edited nearest neighbour rule (ENNR) to the majority classes [75]. ENNR first looks for a specific number of *nearest neighbours* ([44] recommends using 3 neighbours) of the "seed" example, re-classifies it with them and then removes these majority examples, which cause the wrong re-classification. Experiments have shown that NCR outperforms OSS [44].

The best well know informative sampling method is the Synthetic Minority Oversampling Technique (SMOTE) [11]. It is also based on the k nearest neighbourhood, however it exploits it to selectively over-sample the minority class by creating *new synthetic examples*. It considers each minority class example as a "seed" and finds its k-nearest neighbours also from the minority class. Then, according to the user-defined *over-sampling* ratio—o_r, SMOTE randomly selects o_r of these k neighbours and randomly introduces new examples along the lines connecting the seed example with these selected neighbours. It generate artificial examples for both qualitative and quantitative attribute.

Some of the assumptions behind SMOTE could still be questioned. First, using the same over-sampling ratio to all minority examples may be doubtful for some data. Several researchers claim that unsafe examples are more liable to be misclassified, while safe examples located inside the class regions are easier to be learned and do not require such a strong over-sampling. What is more important, SMOTE may over-generalize the minority class as it blindly over-samples the attribute space without considering the distribution of neighbours from the majority class. To overcome such limitations several generalizations of SMOTE have been recently introduced; for reviews see [48, 62]. They usually follow one of the two directions: (1) an integration of standard SMOTE with an extra post-processing step or (2) a modification of an internal sampling strategy.

The first solution is to integrate SMOTE with a post-processing phase including filtering the most harmful examples. For instance, using ENNR after SMOTE performs quite well with tree classifiers [2] and rules [52]. Yet a more elaborated approach is presented in [62], where an additional bagging ensemble is used to identify the most misclassified examples and iteratively remove them if it improves evaluation measures. The other group of more "internal" extensions includes two general solutions. The first generalizations over-sample some types of minority examples only. For instance, in *Borderline-SMOTE* only the borderline examples could be seeds for over-sampling [27]. The other generalizations attempt to modify localizations for introducing the new synthetic examples. In *Safe Level SMOTE* and *LN-SMOTE* the distribution of local sub-areas around the seed example and its selected neighbour are analysed and the new example is generated closer to a safer one [48].

Hybrid methods combine of over-sampling with cleaning difficult examples. Besides a simple integration of SMOTE with either ENN or Tomek links [68] other more complex methods offer sophisticated internal combinations of different operations, e.g. by using evolutionary algorithms to optimize some parameters, as the balancing ratio, combinations of over-sampling versus under-sampling amount, see e.g. [21, 71].

SPIDER is another hybrid method that selectively filters out harmful examples from the majority class and amplifies the difficult minority examples [65]. In the first stage it applies ENNR to distinguish between safe and unsafe examples (depending how k neighbours reclassify the given "seed" example). For the majority class—outliers or the neighbours which misclassify the seed minority example are either removed or relabeled. The remaining unsafe minority examples are additionally replicated depending on the number of majority neighbours.

Note that in all the above mentioned methods k nearest neighbourhood is often calculated with the HVDM metric (*Heterogeneous Value Difference Metric*) [75]. Recall that it aggregates normalized distances for both continuous and qualitative attributes, however it uses the Stanfil and Valtz value difference metric for qualitative attributes.

Many generalizations of ensembles are based on integrating re-sampling to modify contents of training samples in bagging or boosting. For instance, SMOTE-Boost is an integration of SMOTE with classical AdaBoost to focus successive classifiers on the minority class [10]. Another representative is IIvotes, where SPIDER is added to Breiman's importance sampling of bootstraps [6]. Other extensions of bagging re-balance the class distribution inside each bootstrap sample into fully balanced ones, by either simple random over-sampling of the minority examples, or by under-sampling the majority class—for their review and experimental comparison see [7, 19].

2.2 Evaluation Issues

Imbalanced data constitutes a challenge not only when constructing a classifier, but also when evaluating its performance. Indeed, an overall classification accuracy is not the best criterion characterizing performance of a classifier as it is biased toward the majority classes. A good recognition of the minority is more preferred, thus a classifier should be characterized rather by other specialized measures, e.g. by its *sensitivity* and *specificity* for the minority class.

Both these and other similar measures are defined with the confusion matrix for two class only, where typically the class label of the minority class is called positive and the class label of the majority class is negative [29, 35]. Even if data contains more majority classes the classifier performance on these classes are usually aggregated into one negative class.

The *sensitivity* (also called a True-Positive Rate or *Recall* of the minority class) is defined as the ratio of correctly recognized examples from the minority class while the *specificity* is the ratio of correctly excluded examples from the majority classes (in a case of binary classification the specificity of the minority class is the recall of the majority class). More attention is usually given to sensitivity than to specificity [24]. However, in general there is trade-off between these two measures, i.e., improving the sensitivity too much may lead to deterioration of specificity at the same time—see experimental results in [25]. Thus, some measures summarizing both points of view are considered. One of them is *G-mean* [42], calculated as a geometric mean of sensitivity and specificity. Its key idea is to maximise the recognition of each of minority and majority classes while keeping these accuracies balanced. An important, useful property of the G-mean is that it is independent of the distribution of examples between classes. An alternative criterion aggregating precision and recall for the minority class is *F-measure*; for a deeper discussion of its properties see e.g. [29]. Other less frequently used measures are nicely reviewed in [38].

Several authors also use the *ROC* (*Receiver Operating Characteristics*) *curve* analysis in case of scoring classifiers. A ROC curve is a graphical plot of a true positive rate (sensitivity) as a function of false positive rate (1-specificity) along different threshold values characterizing the performance of the studied classifier [35]. The quality of the classifier performance is reflected by the area under a ROC curve (so called AUC measure) [10, 35, 38]. Although AUC is a very popular tool, some researchers have discussed some limitations, e.g. in the case of highly skewed data sets it could lead to an overoptimistic estimation of the algorithm's performance [28]. Thus, other proposals include Precision Recall Curves or other special cost curves (see their review in [13, 29]).

3 Nature of Imbalanced Data

A data set is considered imbalanced when it is characterized by an unequal distribution between classes. N. Japkowicz refers it to the *between-class imbalance* [33]. It is evaluated by the *global class imbalance ratio IR*. Assume that the data set D contains n learning examples assigned to two classes: the minority class MK with N_{min} representatives and the majority class WK having N_{maj} examples. Depending on the literature sources, IR is usually expressed as either N_{maj}/N_{min} or the percentage of N_{min} in the total number of examples n.

There is no unique opinion about the threshold for the degree of such imbalance between the class cardinalities to establish data to be imbalanced. Some researchers have studied the data sets where one class was several times smaller than other classes, while others have considered more severe imbalance ratios as, e.g., with $IR = 10/1$, 100/1 or even greater. Without showing a precise threshold value for this ratio, we repeat after [72] that the problem is associated with lack of data (absolute rarity), i.e. the number of examples in the rare (minority) class is too small to recognize properly the regularities in the data.

Although this description implies binary (two-class) problems, data with many majority classes are often aggregated into one global majority class—which is a case considered in this paper. However, note that some authors also consider multi-class data sets, where imbalances may exist between various classes.

The imbalance of a learning data set can be either *intrinsic* (in the sense that it is a direct result of the nature of the data space) or *extrinsic* (caused by reasons external to the data space). Extrinsic imbalances can be caused by too high costs of acquiring the examples from the minority class, e.g. due to economic or privacy reasons [72] or comes from technical time or storage factors. For instance, He et al. give in [29] examples of learning from continuous balanced data stream where due to technical sporadic interruptions in transmissions of some sub-blocks inside the analyzed stream would become an extrinsic imbalanced data set.

Gary Weiss also discusses problems of data rarity and distinguishes between *relative imbalance* and *absolute rarity*. In the former case, the data set contains too small minority class. However, if it is possible to collect/sample more examples and to increase the total size of data while keeping the same global imbalance ratio, it may happen that the absolute cardinality of the minority class will not be rare anymore and it may be easier to be learned [72].

On the other hand, some studies have shown that for even highly imbalanced data the minority class can be sufficiently accurately learned by all standard algorithms [2]. Examples of such popular UCI benchmark data sets are `new-thyroid` or `vehicle`—see their experimental analysis in [52]. Indeed one can image binary class distributions which could be linearly separated with not so much disturbance from even high imbalances assuming that the minority class does not represent an absolute rarity. In case of a clear separation the minority class boundary could be easily approximated by many algorithms.

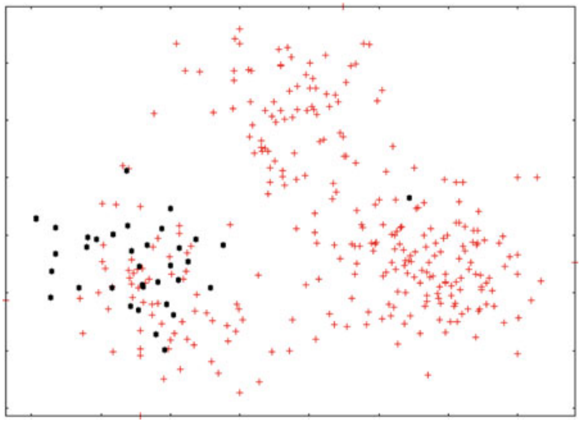


Fig. 1 MDS visualization of class distribution in ecoli imbalanced data

Distributions of real world imbalance data usually are not similar to the above examples. For instance, Napierala in her attempts to visualize imbalanced data with either multi-dimensional scaling or non-linear projections [52] to low dimensional (2 or 3 variables) has showed such distributions as presented in Fig. 1. One can notice that in `ecoli` data both classes are not separated, instead they seriously overlap. The consistent region belonging solely to the minority class is rather very small—most examples lie in a mixed region between the classes. Another observation is presence of small sub-groups of the minority class, having sometimes few instances only.

Furthermore, well known comprehensive experimental studies where many specialized approaches over large collections of imbalanced data show that simply discussing the global imbalance ratio does not sufficiently explain differences of classification performance of these approaches [22, 37, 47, 53, 64].

All these results lead us to conclude that the global imbalance ratio is not the only, and possibly not the main, data factor that hinders deterioration of learning classifiers. As some researchers claims one should rather consider data set *complexity* which should be more influential. *Data complexity* can be understood as the difficult properties distribution of examples from both classes in the attribute space. It is not particularly surprising that it shows a crucial impact on learning, as one could expect that data complexity should affect learning also in balanced domains. However, when data complexity occurs *together* with the class imbalance data difficulty factors, the deterioration of classification performance is amplified and it affects mostly (or sometimes only) the minority class.

The term "data complexity" can comprise different data distribution patterns. Up to now, the researchers have distinguished several *data difficulty factors* which hinder learning in imbalanced domains, such as: decomposing the minority class into rare sub-concepts, overlapping, and presence of outliers, rare instances or noise. We will discuss their role in the next sections.

4 Decomposition of the Minority Class

4.1 Rare Sub-concepts and Small Disjuncts

Nearly all research on data difficulty factors were carried out by experimental studies with synthetic data. The most well known and inspiring studies are research of Nathalie Japkowicz and her co-operators. They focused on *within-class imbalance*, i.e. target concepts (classes) were decomposed into several sub-concepts [33, 37]. To check how much increasing the level of such a decomposition could influence the classification performance, they carried our many experiments with specially generated data. They controlled three parameters: the size of the training set, the imbalance ratio, and so called *degree of concept complexity* (understood as a decomposition of the class into a number of sub-concepts). Two classes were considered—the minority versus the majority class. In their first experiments each data set was generated over a one-dimension interval. Depending on the concept complexity, the input interval was divided into a number of sub-intervals of the same size (up to five), each associated with a different class label. Following similar assumptions, in further studies they generated additional data sets in five-dimensional space, where an occurrence of classes was modeled by separate clusters.

Decision tree (C4.5) and multi layered perceptron neural networks (MLP) were learned from these data sets. The results of their experimental evaluation showed that imbalance ratio did not cause the degradation of classifiers' performance as much as increasing the degree of complexity (the number of sub-intervals). The worst classification results were obtained for the highest decomposition of classes (5 sub-intervals), in particular if they contained too small number of examples. On the other hand, in much larger data, where sub-clusters were represented by a reasonable number of examples, the imbalance ratio alone did not decrease the classification performance as much [37].

According to Japkowicz [33], if such imbalanced sub-concepts contain quite a small number of minority class examples, then the deterioration of classification performance is associated with the problem of so called *small disjuncts*—which was originally introduced by Holte et al. in standard (balanced) learning of symbolic classifiers [32]. Briefly speaking, a classifier learns a concept by generating disjunct forms (e.g. rules of tree) to describe it. Small disjuncts are these parts of the learned classifier which cover a too small number of examples [32, 67, 72]. It has been observed in the empirical studies that small disjuncts contribute to the classification error more than larger disjuncts. In case of fragmented concepts (in particular in the minority class) the presence of small disjunct arises [29]. The impact of small disjuncts was also further studied by other researchers, see e.g. [59, 73]. In particular, additional experiments with applying other classifiers on the artificial data constructed in the similar way as [34] showed that decision trees were the most sensitive to the small disjuncts, then the next was multi layered perceptron, and support vector machines were the less sensitive to them.

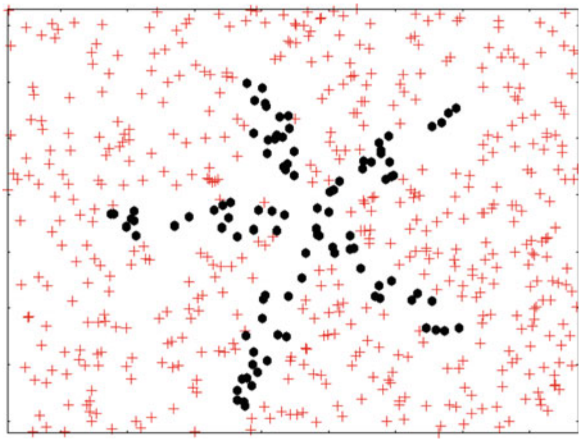


Fig. 2 Clover data set

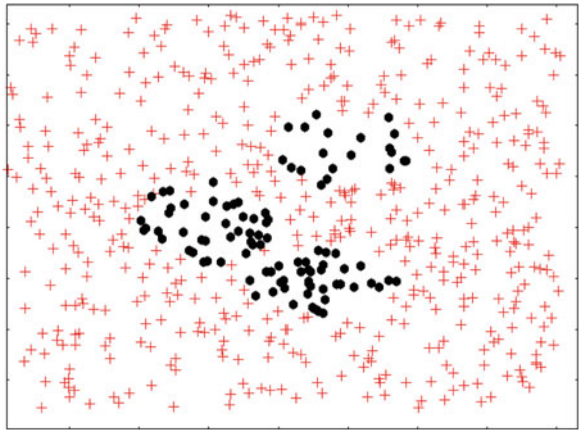


Fig. 3 Paw data set

Stefanowski studied in [64] more complicated decision boundaries in two dimensional, numerical data. First data sets, called `sub-clus`, contained rectangles defining the minority class distributions. All these sub-rectangles are surrounded by the uniformly distributed examples from the majority class. Figure 2 represents the next shape, called a `clover`, a more difficult, non-linear setting, where the minority class resembles a flower with elliptic petals (here 3 sub-concepts—petals). The examples of majority class were uniformly distributed in all the free parts. Similarly to earlier Japkowicz et al. research [37] , the size of data was changed (from 200 to 1200 examples) and the imbalance ratio changed from fully balanced $IR = 1$ till more highly imbalanced $IR = 9$. The minority class was also stepwise decomposed from 2 to 6 sub-parts. Finally, other non-linear shapes of the minority class sub-concepts were presented in `paw` data, see Fig. 3.

Three algorithms: *k*–nearest neighbor (k-NN), decision tree (J4.8)—and rule (JRIP)–based classifiers were considered. Representative results of the sensitivity

Table 1 Sensitivity of k-NN classifier with respect to decomposing the minority class into sub-concepts and changing other parameters of sub-class data

Number of sub-clusters	IR = 5			IR = 9		
	600	400	200	600	400	200
2	0.82	0.8	0.78	0.78	0.76	0.45
3	0.78	0.72	0.70	0.66	0.74	0.25
4	0.75	0.70	0.68	0.64	0.50	0.15
5	0.73	0.68	0.42	0.58	0.45	0.11
6	0.64	0.62	0.36	0.42	0.32	0.10

Table 2 Sensitivity of a tree classifier with respect to decomposing the minority class into sub-concepts and changing imbalance IR

Number of sub-clusters versus IR	600				400			
	3	5	7	9	3	5	7	9
2	0.92	0.92	0.83	0.80	0.94	0.85	0.82	0.80
3	0.90	0.85	0.80	0.78	0.84	0.78	0.72	0.70
4	0.85	0.80	0.78	0.74	0.82	0.75	0.68	0.60
5	0.75	0.35	0.24	0.06	0.14	0.10	0	0
6	0.22	0.10	0	0	0.06	0	0	0

Data size –600 and 400 examples

measure are shown in Table 1 for k-NN classifier and in Table 2 for decision trees. One can notice that while changing the size of the data—larger number 600 and 400 did not influence so much as 200 ones. The highest decrease of evaluation measures (also for G-mean) was observed for increasing the number of sub-regions of the minority class combined with decreasing the size of a data set—for all sizes of data it degraded the performance of a classifier much more than increasing the imbalanced ratio. The tree and rule classifiers showed the similar performance. The degradation of performance was larger if the decision boundary became non-linear even for larger data set. It is illustrated in Table 2 by results for tree classifier and `clover` data. The stepwise growth of the number of sub-regions (from 2 to 6) in `clover` shape decreases much more the sensitivity measure than stepwise increase of the class imbalance ratio (from 3 to 9).

4.2 Dealing with Small Disjuncts

As a consequence of this research special approaches to handle the problem of small disjuncts were proposed in [34, 37]. They are based on specialized over-sampling of the minority class, sometimes also the majority class, with respect to inflate small disjuncts. The most well known proposal is *cluster-based over-sampling* [37]. Its idea is to consider not only the between class imbalance but also the within-class imbalance (imbalance between discovered sub-clusters of each class) and to oversample the data set by improving these two types of imbalances simultaneously. More precisely, the approach workflow is the following:

1. Apply a clustering algorithm to examples from each class separately. In this way, one discovers C_{min} clusters in N_{min} examples from the minority class MK and C_{maj} clusters in N_{maj} examples from the majority class WK.
2. Inside the majority class all the clusters C_{maj}, except the largest one, are randomly oversampled so as to get exactly the same number of examples as inside the largest cluster. In this way the current size of the majority class increases from N_{maj} to Max_{maj}.
3. In the minority class, each cluster is randomly over-sampled until it will contain Max_{maj}/C_{min} examples, where C_{min} is the number of clusters inside the minority class.

As the over-sampled data set will finally contain the same number of examples and all sub-clusters will also be of the same size, the authors claim that no between-class and no within-class imbalances remain inside the transformed data. They successfully applied this approach to several artificial data as well as to 2 real world problems of letter recognition [36] and text classification [57]. In these applications they applied k-means clustering algorithm, although they did not give precise hints how to tune an appropriate k value.

Similarly Borowski [8] considered this pre-processing in text categorization of two larger collection of documents. The first collection was Reuters 21578 and its subset, called MadApte,[1] where 9603 documents constituted a training set (the minority class—`corn`—contained 181 examples) while 3299 ones were used a testing set. The other collection was OHSUMED containing text summaries restricted to subparts from 23 cardiovascular diseases.[2] The training set contained 10,000 summaries (the minority class—`C01 disease`—has 423 documents) while the testing sets was build on 10,000 summaries. In both collections NLP techniques were applied to extract around 5000 terms in a vector space representation. Then features were selected to around a few hundred by using chi-square and entropy gain filters. Tables 3 and 4 summarize the main classification results of using different pre-processing methods with the following classifiers: Naive Bayes (abbreviated as NB), k-nearest neighbour (k-NN), logistic regression (Reg-Log) support vector machines (SVM). For cluster over-sampling we tested 6 values $k = 4, \ldots, 10$—the best values were 6 and 7 depending on data. SMOTE was applied with 5 neighbours and testing oversampling ratios o_r between 100 and 1000 % (with a step 100)—the best ratio was 400. Note that the cluster over-sampling improved both G-mean and F-measure. However, these improvements were not as high as those achieved by using SMOTE.

[1] Reuters data is at http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[2] OSHSUMED available at http://ir.ohsu.edu/ohsumed/ohsumed.html.

Table 3 Applying cluster over-sampling and SMOTE to Reuters data

Method	Classifiers				Evaluation measure
	NB	k-NN	Reg-Log	SVM	
Cluster-oversample	0.42	0.41	0.49	0.45	F
	0.77	0.71	0.77	0.69	G-mean
SMOTE	0.38	0.46	0.47	0.46	F
	0.88	0.83	0.90	0.91	G-mean
No pre-processing	0.0	0.34	0.18	0.4	F
	0.0	0.56	0.33	0.59	G-mean

Table 4 Applying cluster over-sampling and SMOTE to Oshumed data

Method	Classifiers				Evaluation measure
	NB	k-NN	Reg-Log	SVM	
Cluster-oversample	0.46	0.40	0.48	0.43	F
	0.72	0.64	0.71	0.68	G-mean
SMOTE	0.34	0.41	0.47	0.49	F
	0.81	0.77	0.83	0.82	G-mean
No pre-processing	0.13	0.38	0.34	0.46	F
	0.27	0.61	0.51	0.65	G-mean

A quite similar conclusion was reached by another study of Napierala et al. [55] with synthetic data sets—`subclass`, `clover` and `paw`—which were affected by different amounts of disturbance (increasing amount of overlapping and rare examples—this type of examples is further defined in Sect. 6.1). The representative results are presented in Table 5 where base denotes using a classifier without any pre-processing, RO is a simple random over-sampling, CO—cluster over-sampling, NCR—nearest cleaning rule, and the last column refers to SMOTE. While analyzing these results one can notice that cluster over-sampling is competitive with other methods for data sets containing the minority class without any perturbations. Then, the more complex, overlapped and affected shapes of the minority class sub-parts, the better are other pre-processing methods as SMOTE and SPIDER.

Yet another approach to deal with the above-mentioned within class decomposition was presented in [26]. Gumkowski and Stefanowski proposed to use a two phase approach including: (1) clustering and (2) constructing a hierarchical classifiers. More precisely,

1. Use a clustering algorithm to identify sub-concepts of the minority class.
2. Construct Voronoi diagram sub-regions around centroids of the identified minority class clusters; Assign also majority class examples to these sub-regions following the distance to the nearest centroid of the minority class cluster.

Table 5 G-mean for synthetic data sets with varying degrees of the disturbance ratio

Data set	Pre-processing method					
	Base	RO	CO	NCR	SPIDER	SMOTE
subclus-0	0.937	0.937	0.948	0.925	0.929	0.938
subclus-30	0.733	0.724	0.724	0.702	0.715	0.712
subclus-50	0.559	0.565	0.602	0.664	0.621	0.704
subclus-70	0.407	0.442	0.469	0.596	0.578	0.541
clover-0	0.739	0.742	0.761	0.778	0.791	0.738
clover-30	0.636	0.637	0.651	0.722	0.676	0.665
clover-50	0.506	0.554	0.549	0.696	0.607	0.601
clover-70	0.418	0.465	0.489	0.658	0.568	0.571
paw-0	0.904	0.913	0.918	0.918	0.902	0.968
paw-30	0.763	0.776	0.771	0.785	0.778	0.833
paw-50	0.657	0.686	0.686	0.752	0.712	0.786
paw-70	0.508	0.582	0.569	0.718	0.651	0.718

Table 6 G-mean performance of the hierarchical classifiers with cluster analysis (HC) against a standard decision tree (J4.8)

Data	Classifier	Sensitivity	F	G-mean
paw-0	treeJ48	0.855	0.968	0.713
	HC	0.940	0.975	0.844
paw-separ	treeJ48	0.98	0.925	0.739
	HC	0.961	0.946	0.864
paw-overlap	treeJ48	0.0	0.0	0.0
	HC	0.741	0.81	0.614
paw-outliers	treeJ48	0.0	0.0	0.0
	HC	0.86	0.89	0.729

3. Learn separate classifiers from learning examples (from both classes) located in each sub-region.
4. Built the arbiter for the set of classifiers—i.e. for a new instance, find to which Voronoi region it belongs and use its classifier to make a final decision.

This approach was implemented in WEKA with X-means clustering algorithm and J4.8 decision trees and its resulting classifier will be further abbreviated as *HC*. X-means is a kind of wrapper around running k-means with different k. The resulting clustering is chosen with respect to optimizing BIC criterion [51].

In Table 6 we show results of using this approach with J4.8 trees over several versions of the synthetic data set `paw`. We used it as it models three different sub-concepts inside the minority class (see its idea in Fig. 3). The first data, called `paw-0` is just a version illustrated in this figure without any disturbance. In this construction two sub-concepts are quite close to each other, so may mislead the clustering algorithm (X-means has a tendency to propose 2 clusters instead of three clusters). Therefore, we constructed a version with more separated clusters (moving clusters away)—this is called `paw-separ`. Then, we additionally disturbed minority class shapes by introducing overlapping (`paw-overlap`) and moving more minority examples inside the majority class as outliers.

In case of these synthetic data sets `paw`, where sub-parts are relatively well separated, this algorithm can divide the space into three sub-areas and the hierarchical classifier *HC* improves slightly the sensitivity and other measures comparing to using a single, standard tree. The improvements are a bit higher for `paw-0`, with more difficult separation. For more disturbed data `paws` with overlapping and outliers the standard trees deteriorates its performance while the *HC* classifier maintains its good performance—although values of evaluation measures are smaller than in cleaner shapes. However, we can conclude that in all cases the proposed approach improves evaluation measures.

4.3 Open Issues

Although the idea of identifying and dealing with small disjuncts sounds quite appealing, several open issues remain critical if one needs to analyse real-world data sets. Note that most of the above discussed studies have been carried out with special synthetic data while for real ones the underlying distribution of the minority class is unknown and it is not easy to approximate (or even to guess) the possible number and structure of sub-concepts.

Up to now most researchers have used clustering algorithms to find these sub-concepts. Other, quite rare studies concern analyzing classification or association rules, mainly their evaluation measures as coverage or similar ones, to distinguish between small and large disjuncts.

For clustering nearly all authors applied k-means algorithm. The main problem is to tune k number of searched clusters. However, other problems include dealing with non-spherical, complex shapes of clusters, an influence of overlapping, noise or outliers. It is also not obvious which clustering evaluation criteria should be considered as the most common ones were proposed for standard unsupervised framework [49]. Here, one deals with at least partly supervised and imbalanced case where one has to distinguish between minority and majority examples inside clusters. Even if clustering algorithms are applied separately to each class the algorithm may improperly agglomerate smaller sub-parts of the minority into too large ones (see experiences with paw data in [26]).

Tuning clustering algorithm parameters in the wrapper framework is also non-trivial. First, it refers to choosing an appropriate quality criterion. Some authors propose to consider tuning clustering together with learning the final classifier and evaluate the integrated approach with special imbalance measures (as e.g. G-mean, AUC). To avoid possible over-fitting it is necessary to use an extra validation set or an internal cross validation inside the training set. This was a case in experimental studies as [8, 61]. However, one should take into account that the data set may be highly imbalanced and it may be difficult, or even impossible, to select a sufficient number of minority examples inside learning, validation and test parts. Perhaps new solutions of partly informed bootstrap sampling could be developed. One should also remember that scanning too many *k* may be too time consuming or even not realistic.

Nevertheless, k-means may not be the best algorithm to be used in this context, in particular for more complex distributions of the minority class which we will discuss in further sections. Besides non-linear and complex decision shapes of clusters, overlapping many minority examples could be either singletons like outliers or rare cases (a kind of pairs or triples). Additional experiments with real data sets showed that approaches such as clustering or building hierarchical classifiers are not effective for such difficult data sets [26, 55]. Moving toward density based clustering algorithms is one of the solutions. They can distinguish between core instances (creating clusters) and noisy ones (referring to outliers or rare cases). However tuning parameters of DBSCAN or OPTICS is also not an easy task even in a classical unsupervised version [17]. The current heuristics do not take into account a distinction between minority and majority examples but treat them in the same unsupervised way. Some recent proposals of working with DBSCAN try to look for new heuristics [58]. However, we think that it is necessary to develop a new family of *semi-supervised density algorithms* which take into account labels of examples while constructing neighbour clusters. Finally as imbalanced data sets usually contain minority class outliers new approaches for their detection are still necessary.

5 Overlapping Between Minority and Majority Classes

Researchers also study different difficulty factors characterizing imbalanced data. An *overlapping* between minority and majority classes is one of them. Although many previous studies in classical machine learning have shown that overlapping of different classes deteriorates the total classification accuracy, its influences on the minority class is thoroughly examined. As the minority class is under-represented in the data set, it will be more likely under-represented also in the overlapping region. As a result, the algorithms may shift the decision boundary too close to the minority class, in the worst case treating the whole overlapping area as belonging to the majority class definition.

5.1 Experimental Studies

Prati et al. started more systematic studies on the role of overlapping [60]. They generated artificial data sets where the minority and the majority class were represented by two clusters in five dimensional space (examples where generated around centroids following the Gaussian distribution). Two parameters were investigated:

the imbalance ratio, and the distance between centroids—so classes could be moved from clear separation to high overlapping. For the C4.5 classifier they showed that increasing the overlapping ratio was more responsible for decreasing AUC results than decreasing cardinality of the minority class.

Then, an influence of increasing overlapping was more precisely examined in [22]. Garcia et al. generated two-dimensional data sets with two classes separated by a line orthogonal to one of the axis. They assumed a fixed size of data and changed the overlapping amount for a given imbalance ratio and vice versa. Results of experiments with 6 different classifiers showed that increasing overlapping degraded their performance more (with respect to minority class) than changing the imbalance ratio. Moreover, in the other experiments they fixed the amount of overlapping and changed the distribution of the minority examples by increasing their number in the overlapping area. Again the results confirmed that increasing the local imbalance ratio and the size of the overlapping area were more influential than changing the overall imbalance ratio. However, these factors influenced performance of particular classifiers in a different way. For instance *k*—nearest neighbor classifier was the most sensitive to changes in the local imbalance region. Naive Bayes, MLP and J4.8 worked better in the dense overlapping region. These conclusions have been later verified in additional experiments (focusing on performance of k-NN and other evaluation measures), see [23]. One of their conclusions was that when overlapping regions increased, the more local classifiers—like k-NN with smaller values of *k*—performed better with recognition of the minority class.

The other study in [14] focused on the effects of overlapping and class imbalance on support vector machines (SVM). The authors showed that when the overlap level was high, it was unlikely that collecting more training data would produce a more accurate classifier. They also observed that the performance of SVM decreased gradually with the increasing imbalance ratio and overlapping, and that there was a sudden drop when the imbalance ratio equaled to 20 % and the overlapping level exceeded 60 %, regardless of the training set size.

Prati et al. have recently come back to studying the overlapping in class imbalance [3]. Comparing to their previous work [60] they investigated the usefulness of five different re-sampling methods on the same difficult artificial data sets: popular random-over sampling, random under-sampling, Nearest Cleaning Rule (NCR) [44], SMOTE and SMOTE + ENN [11]. Their main conclusion was that appropriate balancing of training data usually led to a performance improvement of C4.5 classifiers for highly imbalanced data sets with highly overlapped classes. However, the improvements depend on the particular pre-processing method and the overlapping degree. For the highest degree of overlapping it was not clear which method was the best (NCR worked there quite well). Results for other overlapping degrees showed that over-sampling methods in general, and SMOTE-based methods in particular, were more effective than under-sampling. Then, the data cleaning step used in the SMOTE + ENN seemed to be especially suitable in situations having a higher degree of overlapping.

Finally, we come back to our studies [39, 64] where the effect of overlapping was studied together with other factors such as decomposition of the minority class into

Table 7 Influence of overlapping on the sensitivity of the tree classifier learned from subclass data

Number of sub-clusters	IR = 5			IR = 9		
	0 %	10 %	20 %	0 %	10 %	20 %
3	0.96	0.91	0.85	0.94	0.9	0.75
4	0.96	0.89	0.78	0.94	0.87	0.74
5	0.96	0.87	0.76	0.90	0.81	0.66
6	0.94	0.84	0.74	0.88	0.68	0.38

Overlapping is expressed by % of borderline examples from the minority class. Total number of examples –800

smaller sub-concepts and more complicated non-linear borders. The k-NN, rules (MODLEM [63]) and J4.8 decision tree classifier were applied to a collection of specially generated artificial data sets sub-class, clover (described in the previous section). Table 7 shows influence of stepwise increase of the amount overlapping on the tree classifier. The degree of overlapping is measured as a percentage of the size of the minority class. It was observed that stepwise increase of overlapping more strongly decrease the sensitivity. For instance, let us analyse the first column (%)—the sensitivity changes from 0.96 to 0.94. While for any of the number of sub-clusters the sensitivity decreases in range of nearly 0.2 (see, e.g. 4 sub-clusters, the sensitivity decreases from 0.96 to 0.78). The similar tendency can be observed for rule and k-NN classifiers.

The influence of overlapping on specialized pre-processing was also studied in [55]. The tree and rule classifiers (J4.8 and MODLEM) were integrated with standard random over-sampling, cluster over-sampling, nearest cleaning rule and SPIDER. All these methods were applied to artificial data sets as `sub-clus`, `clover` and also more complicated versions of `paw` data. The results clearly showed that all methods of pre-processing improved the sensitivity of both classifiers. However, simpler random over-sampling and cluster over-sampling performed comparably on all non-disturbed data sets. While on more difficult sets (disturbance over 30 %) both NCR and SPIDER methods were clearly better than there over-sampling methods.

5.2 Detecting Overlapping in Real World Data Sets

Note that the data difficulty factors, as overlapping, were examined using mostly artificial data sets [22, 60, 64], in which the data distribution was given a priori and the degree of each factor could be precisely controlled by augmenting or diminishing the degree of overlapping [22, 60] as well as the number and cardinality of small disjuncts [36, 37]. Moreover, the data sets were usually two-dimensional.

The difficult issue is to analyse data factors in real-world imbalanced data sets where the *natural* underlying distribution of minority examples is unknown and has to discovered or at least approximated. Although some researchers published wide comprehensive experimental studies with several popular UCI benchmark data—see e.g. [2, 19, 40, 70], nearly all of them are just comparative experiments of using different standard classifiers [47, 59, 70], ensembles [7, 19] or pre-processing methods [2]. The authors have mainly tried to identify general differences of studied algorithms, quite often without a deeper analysis of hidden data characteristics, or referred to averaged values of these data factors which were easier to be calculated as the total number of minority examples or the global imbalance ratio.

There is not so much research on direct evaluation of overlapping in the real world data sets. For example, in [14] (concerning the effects of overlapping and imbalance on the SVM classifier), the authors propose to estimate the degree of overlapping in real-world data sets by measuring a number of support vectors which can be removed from the classifier without deteriorating the classification accuracy. In the next chapter we will present a simpler and intuitive method based on analyzing local characteristics of minority examples.

6 Types of Minority Examples with Respect to Their Local Characteristics

6.1 Motivations

The first paper discussing different types of minority examples is [42] where Kubat and Matwin have distinguished between safe, borderline and noisy examples. *Borderline examples* are located in the area surrounding class boundaries, where the minority and majority classes overlap. However, they are not only located in the overlapping zone (discussed in the previous section) as they could also be difficult examples located in specific sub-areas near highly non-linear decision boundaries. *Safe examples* are placed in relatively homogeneous areas with respect to the class label. So, they are inside typical clear parts of target concepts, which located further from the decision boundary between classes. By *noisy examples* they understand individuals from one class occurring in safe areas of the other class. However, authors applied this term to majority class individuals inside the minority class and proposed to remove them from the training set [42].

Here we share these authors' opinions saying that as the minority class is often under-represented in the data, one should be careful with the similar treatment of the singletons from the minority class and rather not recognizing them as noise. Moreover, it is worth to stress that the typical understanding of noisy examples in machine learning corresponds to a kind of data imperfections or errors (see e.g. [20, 62, 69]) which come from either random errors in descriptions of examples or from an inappropriate description language. Researchers usually distinguish between class noise (errors with assigning a correct class label to a given example) or attribute noise (erroneous attribute values which could lead to wrong decisions, in particular, if such an example is located too close to decision boundaries) [9]. The typical approaches to deal with such noisy examples include: (1) identification of suspicious examples and eliminating or correcting them (e.g., by using edited approaches for k-nearest neighbour classifiers) or (2) omitting them during the learning phase to solve overfitting phenomena (e.g., by pruning in decision trees or rules). These approaches may improve the total classification accuracy in the standard learning perspective, see e.g. [9, 20].

However, the role of noisy examples in imbalanced data has not been deeply studied yet. Some authors randomly injected changes of class labels or attribute values to noise free data [1, 62, 69]. In such a way in [1, 69] effectiveness of standard techniques for handling class noise was evaluated. These two independent experimental results showed that all learning algorithms were sensitive to noise in the minority examples, however some of them, such as Naive Bayes and k—nearest neighbor classifiers, were often more robust than more complex learners, such as support vector machines or Random Forests. In more recent our studies [62], the authors introduced both class noise and attribute noise, by either changing the class label or the attribute values, respectively. The comparison concerned the SMOTE preprocessing method and its several extensions. It showed that SMOTE was sensitive to the noisy data and its extensions which also clean noise introduced by SMOTE, were necessary. In particular, the new proposed specialized noise filter added as post-processing to SMOTE, called SMOTE-IPF, can deal with some of these noisy examples [62].

Napierala and Stefanowski in their papers [52–55] claimed that one should be very careful with directly transferring standard methods for dealing with noise to difficult minority class examples, as it may lead to removal or relabel too high number of minority examples, or to prune too many elements of classifiers mainly for the minority class. This claim is also consistent with research of Koshgoftar et al. [69] which also stated that in the class imbalance setting, using standard approaches for handling noise “can be catastrophic”. The study in [9] also showed that when there is an abundance of data, it is better to detect properly “bad data” at the expense of throwing away “good data”, while in case when the data are rare, more conservative filters are better.

What is even more important—the noisy examples are often misclassified with singletons playing a role of *outliers*. Note that the outlier is just an untypical example not coming from erroneous measurements. As the minority class can be under-represented in the data, the minority class singletons located in the majority class areas can be outliers, representing a rare but valid sub-concept of which no other representatives could be collected for training. A quite similar opinion was expressed e.g. in [42], where the authors suggested that minority examples should not be removed as they are too rare to be wasted, even considering the danger that some of them are noisy. In [76], which concerns the detection of noise in balanced data sets, the authors suggest to be cautious when performing automatic noise correction, as it may lead to ignoring outliers which is “questionable, especially when the users are very serious with their data". In our opinion, the minority class examples conform to this case.

We claim that the minority and majority distant examples should be treated in a different way. Majority examples located inside the minority class regions are more likely to be a true noise and they could be candidates for removing or relabeling. In general, noisy majority examples are undesired as they can cause additional fragmentation of the minority class and can increase the difficulties in learning its definition. On the other hand, minority examples considered as outliers should be rather kept in the learning set and properly treated by next pre-processing methods or specialized algorithms for imbalanced data.

Moreover, it is worth to distinguish yet another type of so-called *rare examples*. These are pairs or triples of minority class examples, located inside the majority class region, which are distant from the decision boundary so they are not borderline examples, and at the same time are not pure singletons. The role of these examples has been preliminarily studied in the experiments with special artificial data sets [55, 64]. It has been shown that rare examples significantly degraded the performance of classifiers. Additionally, various pre-processing methods performed differently on such rare examples. Finally, works on graphical visualizations of real-world imbalanced data sets [53] have shown existence of such types of examples. The reader can also analyse Fig. 1 where the minority class contains mainly unsafe examples: many borderline, pairs or triples of rare small "islands" and many outliers.

Napierala and Stefanowski in their earlier research [53, 55] claimed that many of considered data difficulty factors could be linked to the distinguishing the following types of examples forming the minority class distribution:

- safe examples
- borderline examples
- rare examples
- outliers

They also claimed that distinguishing these types of examples can be useful to focus attention on difficulties of the minority class distributions, to support interpretations of differences in the performance of classifiers or specialized methods applied to imbalanced data as well as to develop new specialized algorithms. In the next subsection we will briefly discuss some of these issues.

6.2 Identification of Example Types

Distinguishing four types of examples refers to most of previously discussed data difficulty factors. If the minority class distribution will contain mainly unsafe examples, it could indicate that the minority class does not constitute a whole concept but is affected by different disturbances. Although one cannot directly discover subconcepts, it is possible to indirectly show possible decomposition. A larger number of borderline examples will directly approximate overlapping difficulty factors. Furthermore, rare examples and outliers also express data difficulty discussed in the previous sub-section. Finally, it should be stressed that authors of related works focus rather on studying single data factors and usually do not consider several data factors occurring together. What is even more important to notice, they usually carried out their experiments with artificially generated data, where given perturbations were introduced to assumed data distribution, and rarely attempt to transfer such studies to real world methods.

Therefore, while considering our distinguishing of four types of examples, the research open issue is—how does one can automatically and possibly simply identify these example type in real world data sets (with unknown underlying class distributions).

Note that the visualisation projection methods—discussed in [52]—could confirm the occurrence of different types of examples in some real-world data sets but they cannot be directly applied in the real-world settings. First of all, they cannot be used for very large data sets, as the visualisation of thousands of points would be difficult to read. Secondly, the projection to two dimensions may not always be feasible, as the data set may be intrinsically characterized by more dimensions.

Furthermore, as we attempt to stress in earlier sections, it is practically easy to directly measure only the simplest data characteristics as the global imbalanced ratio, data size, etc. while other more influential data factors are rather difficult to precisely estimate in real world, not trivial data sets. Some of already proposed methods may rather very roughly indicate the presence of the given data factors. For instance, in [14] (concerning the effects of overlapping and imbalance on the support vector machine classifier), the authors proposed to approximate the possible amount of overlapping in real-world data sets by measuring a number of support vectors which can be removed from the classifier without deteriorating the classification accuracy. Other methods for identification of unsafe or noisy examples are based on an extensive using cross-validated ensembles, bagging and boosting. However, their parameters are not easy to tune. Moreover, not all instances misclassified by ensembles may be noisy examples as some of them could be rather difficult but valid examples.

Therefore, Napierala and Stefanowski have looked for new simple techniques which should more directly identify the difficult types of example distributions in imbalanced data. Moreover, they could be more intuitive for user with respect to principles and rules of their parametrization.

The proposed method origins from the hypotheses [53] on role of the mutual positions of the learning examples in the attribute space and the idea of assessing the type of example by analyzing class labels of the other examples in its *local neighbourhood*. By a term local we understand that one should focus on the processing characteristics of the nearest examples due to the possible sparse decomposition of the minority class into rather rare sub-concepts with non-linear decision boundaries. Considering a larger size of the neighbourhood may not reflect the underlying distribution of the minority class.

In general, such a neighbourhood of the minority class example could be modeled in different ways. In further considerations we will use an analysis of the class labels among *k-nearest neighbours* [52, 53]. An alternative approach to model the local neighbourhood with *kernel functions* has been recently presented in [52]—however, its experimental evaluation has given similar conclusions as to data characteristics.

Constructing the k—neighbourhood involves decisions on choosing the value of k and the *distance function*. In our previous considerations we have followed results of analyzing different distance metrics [46] and chose the HVDM metric (*Heterogeneous Value Difference Metric*) [75]. Its main advantage for mixed attributes is that it aggregates normalized distances for qualitative and quantitative attributes. In particular, comparing to other metrics HVDM provides more appropriate handling of qualitative attributes as instead of simple value matching, as it makes use of the class information to compute attribute value conditional probabilities by using a Stanfil and Valtz value difference metric for nominal attributes [75]. Tuning k value should be done more carefully. In general, different values may be considered depending on the data set characteristic. Values smaller than 5, e.g. $k = 3$, may poorly distinguish the nature of examples, especially if one wants to assign them to four types. Too high values, on the other hand, would be inconsistent with the assumption of the locality of the method and not useful while dealing with complex, non-linear and fragmented distributions of the minority class. In this paper we do not solve the problem of an automatic tuning this value with respect to complexity of the minority class distribution and its difficulty factors, leaving it for future research.

Experiments from [52] over many UCI data sets have showed that choosing k = 5, 7, 9 and 11 values has led to quite similar categorizations of data with respect to proportions of the minority class types. Below we will show assigning types minority class for the smallest k values.

Depending on the number of examples from the majority class in the local neighbourhood of the given minority class example, we can evaluate whether this example could be safe or unsafe (difficult) to be learned. If all, or nearly all, its neighbours belong the minority class, this example is treated as the safe example, otherwise it is one of unsafe types. For instance, in case of $k = 5$ the type of example x is defined as follows:

- if 5 or 4 of its neighbours belong to the same class as x, it is treated as a safe example;
- if the numbers of neighbours from both classes are similar (proportions 3:2 or 2:3)—it is a borderline example;
- if it has only one neighbour with the same label (1:4) it is a rare example;
- if all neighbours come from the opposite class (0:5)—it is an outlier.

Similar interpretations can be extended for larger values of k. For instance, in case of $k = 7$ and the neighbourhood distribution 7:0 or 6:1 or 5:2—a safe example; 4:3 or 3:4—a borderline example; again the number of neighbours from both classes are approximately the same; 2:5 or 1:6—a rare example; and 0:7—an outlier. Such an interpretation can be extended for larger neighbourhoods and even tuning bandwidth in kernels—see such an analysis in [52].

The analysis of this neighbourhood has been applied in experiments with UCI imbalanced real-world data sets [52, 53]. The results of labeling types of minority class examples are presented in Table 8. Note that many data sets contain rather a small number of safe minority examples. The exceptions are three data sets composed of almost only safe examples: `flags, breast-w, car`. On the other hand, there are data sets such as `cleveland, balance-scale` or `solar-flare`, which do not contain any safe examples. We carried out a similar neighbourhood analysis for the majority classes and made a contrary observation—nearly all data sets contain mainly safe majority examples (e.g. `yeast`: 98.5 %, `ecoli`: 91.7 %) and sometimes a limited number of borderline examples (e.g. `balance-scale`: 84.5 % safe and 15.6 % borderline examples). What is even more important, nearly all data sets do not contain any majority outliers and at most 2 % of rare examples. These results show that outliers and rare examples can constitute an important part of the minority class—there are some data sets where they even prevail in the minority class. Therefore, one should be cautious with considering all of them as noise and applying noise-handling methods such as relabeling or removing these examples from the learning set.

6.3 Influence of Example Types on Classification Performance

The results of labeling the minority class examples can also be used to categorize data sets. depending on the dominating type of examples from the minority class. Only in `abdominal-pain, acl, new-thyroid and vehicle` data sets, safe minority examples prevail. Therefore, we can treat these 4 data sets as representatives of *safe* data sets. In the next category the borderline examples dominate in the distribution of the minority class. As could be observed in Table 8, even in data sets with clean borders a considerable amount of examples (up to 36 %) can be labeled as borderline ones. So, the percentage of borderline examples must be even higher to represent some overlapping between classes. We could treat a data set as a *borderline* data set if it contains more than 50 % of borderline examples—for instance these are `credit-g, ecoli, haberman, hepatitis`. Additional data sets—as `car` and `scrotal-pain`—are located somewhere between safe and borderline categories. As the amount of safe examples is too low, they are mostly inside the borderline category. Then, several data sets contain many rare examples. Although they are not as numerous as borderline examples, they constitute even 20–30% of the minority class. The rare category includes `haberman` (also assigned to borderline category), `cmc, breast-cancer, cleveland, glass, hsv` and `abalone` data sets, which have at least 20 % of rare examples. Other data sets contain less than 10 % of these examples. Finally, some data sets contain a relatively high number of outlier examples—sometimes more than a half of the whole minority class. We can assign the data set to outlier category if more than 20 % of examples are labeled as outliers.

In previous studies [52, 53] we compared different learning algorithms and shown that distinguishing these data characteristics is co-related with differentiating differences in the performance of classifiers. First, for the safe data nearly all compared single classifiers (SVM, RBF, k-NN, J4.8 decision trees or PART rules) perform quite well with respect to sensitivity, F-measure or G-mean. The larger differentiation of

Table 8 Labeling minority examples expressed as a percentage of each type of examples occurring in this class

Data set	Safe	Border	Rare	Outlier
abdominal_pain	61.39	23.76	6.93	7.92
balance-scale	0.00	0.00	8.16	91.84
breast-cancer	21.18	38.82	27.06	12.94
breast-w	91.29	7.88	0.00	0.83
bupa	20.69	76.55	0.00	2.76
car	47.83	47.83	0.00	4.35
cleveland	0.00	45.71	8.57	45.71
cmc	13.81	53.15	14.41	18.62
credit-g	15.67	61.33	12.33	10.67
ecoli	28.57	54.29	2.86	14.29
flags	100.00	0.00	0.00	0.00
haberman	4.94	61.73	18.52	14.81
hepatitis	18.75	62.50	6.25	12.50
hsv	0.00	0.00	28.57	71.43
ionosphere	44.44	30.95	11.90	12.70
new-thyroid	68.57	31.43	0.00	0.00
pima	29.85	56.34	5.22	8.58
postoperative	0.00	41.67	29.17	29.17
scrotal_pain	50.85	33.90	10.17	5.08
solar-flareF	2.33	41.86	16.28	39.53
transfusion	18.54	47.19	11.24	23.03
vehicle	74.37	24.62	0.00	1.01
yeast-ME2	5.88	47.06	7.84	39.22

classifiers occurs for more unsafe data sets. For instance, SVM and RBF classifiers work much better for safe category, while rare or outlier data strongly deteriorate their classification performance. Rare and especially outlier examples are extremely difficult to recognize. PART, J48 and sometimes 1NN may classify them but at a very low level. On the other hand, SVM and RBF fail to classify minority examples in these data sets.

Similar analysis has been carried out for the most representative pre-processing approaches, showing that the competence area of each method depends on the data difficulty level, based on the types of minority class examples [56]. Again in the case of safe data there are no significant differences between the compared methods—even random over-sampling works quite accurate. However, for borderline data sets Nearest Cleaning Rules performs best. On the other hand, SMOTE [11] and SPIDER [65], which can add new examples to the data, have proved to be more suitable for rare and outlier data sets.

For more details on the competence of each studied single classifier and pre-processing methods see [52]. Moreover, our results often confirm the results of the related works conducted on artificial data sets, see [2, 23, 55].

Finally, yet another analysis for different generalizations of bagging ensembles specialized for class imbalances, have been carried out in our recent papers [5, 7]. For safe data sets nearly all bagging extensions for imbalanced data achieve similar high performance. The strong differences between classifiers occur for the most difficult data distributions with a limited number of safe minority examples. Furthermore, the best improvements of all evaluation measures for Roughly Balanced Bagging and Nearest Balanced Bagging are observed for the most unsafe data sets with many rare examples and outliers [5].

7 Final Remarks and Open Research Challenges

This paper concerns problems of learning classifiers from imbalanced data. Although many specialized methods have been introduced, it is still a challenging problem. We claim that besides developing new algorithms for improving classifiers, it is more interesting to ask more general research questions on the nature of the class imbalance problem, properties of an underlying distribution of the minority class in data, and its influence on performance of various classifiers and pre-processing methods.

The main aim of this study is to discuss the data difficulty factors which correspond to sources of difficulties in recognizing the minority class. Following the literature survey and own studies we have focused our attention on the following factors:

- decomposition of the minority class into rare sub-concepts,
- overlapping of classes and borderline examples,
- distinguishing different types of the minority examples.

For each difficulty factor we have discussed its influence of classification performance and details of its practical identification in real data sets. The main lesson from various experiments is that these factors are more influential than the global imbalance ratio or the absolute size of the minority class which have been more often considered in the related literature up to now.

Our experiments with synthetics data have clearly showed that increasing data complexity (understood as decomposition of the minority class into many subparts) decreased evaluation measures more than changing the imbalance ratio or the absolute size of the class. We have also showed that combining the minority class decomposition with non-linear decision boundaries and overlapping makes the learning task extremely difficult. However, as it has been discussed and showed on several illustrative examples, identification of sub-clusters (corresponding to small disjuncts) in real world data, e.g. by clustering algorithms, is still an open research challenge. In particular, it is not obvious how to tune algorithm parameters (e.g. a number of expected clusters in k-mean) and to deal with complex shapes or outliers. We think that developing a new kind of a semi-supervised density based algorithm (where it is necessary to deal with presence of minority versus majority examples inside clusters) could be a promising research direction. Similar limitations are manifested by current methods for identification of overlapping minority and majority class distributions.

The other novel contributions are distinguishing different types of minority examples and proposing a new method for their identification in real world data sets. This identification method is based on analyzing class distribution inside the local *k*-neighbourhood of the minority examples. It can also approximate many discussed data difficulty factors, except discovering small disjuncts. Its experimental evaluation has led us to several novel observations with respect to earlier studies on imbalanced data. First, analyzing types of examples in many UCI imbalanced data sets has showed that safe examples are uncommon in most of the imbalanced data. They rather contain all types of examples, but in different proportions. Depending on the dominating type of identified minority examples, the considered data sets could be categorized as: safe, border, rare or outlier. Borderline examples appear in most of the data sets and often constitute more than a half of the minority class. We could also observe that rare and outlier examples are not only extremely difficult for most of the learning methods, but they are often quite numerous in the imbalanced data sets.

Our other comparative experiments have showed that the classifier performance could be related to the above mentioned categories of data. First, for the safe data nearly all compared single classifiers perform quite well. The larger differentiation occurs for more unsafe data set. For instance, support vector machines and RBF neural networks work much better for safe data category, while rare or outlier data strongly deteriorate their classification performance. On the other hand, unpruned decision trees and k-NN classifiers work better for more unsafe data sets. Similar analysis has been carried out for the most representative pre-processing approaches, showing that the competence area of each method also depends on the data difficulty level; For more details see [52]. The other experiments for different generalizations of bagging ensembles for class imbalances, have been carried out in the recent paper [7].

We also claim that the appropriate treatment of these factors, in particular types of minority example, within new proposals of either pre-processing or classifiers, should lead to improving their classification performance. Although it is not inside the scope of this paper, we mention that such research has already been undertaken and resulted in proposing: informed pre-processing method LN-SMOTE [48], rule induction algorithm BRACID [54] and nearest neighbour generalization of bagging, called NBBag [5].

On the other hand, several topics still remain open issues for future research. Besides already mentioned semi-supervised clustering for detecting small disjuncts, one could look for a more flexible method of tuning *k* in the local neighborhood method for identification of types of examples with respect to the given data set; studying differences between outliers and real noise; detecting singleton examples in empty spaces (which is an absolute rarity different to the situation of single examples surrounded by *k*-neighbours from opposite classes), developing a new method for dealing with such examples, re-considering *k*-neighbourhood methods in highly dimensional spaces, studying different over-sampling with respect to identified different characteristics of sub-areas of data. Finally, it is worth to consider mutli-class imbalanced problems, where at least two smaller classes are particularly interesting to experts and they prefer to improve their recognition separately and do not allow to aggregate them together. Although some authors have already attempted to decompose this problem into one-against all or pairwise coupling classifiers, we think it would be more beneficial to look for another framework with unequal costs of misclassifications between classes.

Acknowledgments The research was funded by the the Polish National Science Center, grant no. DEC-2013/11/B/ST6/00963. Close co-operation with Krystyna Napierala in research on types of examples is also acknowledged.

References

1. Anyfantis D, Karagiannopoulos M, Kotsiantis S, Pintelas P (2007) Robustness of learning techniques in handling class noise in imbalanced datasets. In: Proceedings of the IFIP conference on artificial intelligence applications and innovations, pp 21–28
2. Batista G, Prati R, Monard M (2004) A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl 6(1):20–29
3. Batista G, Prati R, Monard M (2005) Balancing strategies and class overlapping. In: Proceedings of the IDA 2005, LNCS vol 3646, pp 24–35, Springer
4. Bishop Ch (2006) Pattern recognition and machine learning. Information science and statistics. Springer, New York
5. Błaszczyński J, Stefanowski J (2015) Neighbourhood sampling in bagging for imbalanced data. Neurocomputing 150(Part B):529–542
6. Błaszczyński J, Deckert M, Stefanowski J, Wilk Sz (2010) Integrating selective pre-processing of imbalanced data with Ivotes ensemble. In: Proceedings of the 7th international conference RSCTC 2010, LNAI vol 6086, pp 148–157, Springer
7. Błaszczyński J, Stefanowski J, Idkowiak L (2013) Extending bagging for imbalanced data. In: Proceedings of the 8th CORES 2013, Springer Series on Advances in Intelligent Systems and Computing, vol 226, pp 269–278
8. Borowski J (2014) Constructing data representations and classification of imbalanced text documents. Master Thesis, Poznan University of Technology (supervised by Stefanowski J.)
9. Brodley CE, Friedl M (1999) A: Identifying mislabeled training data. J Artif Intell Res 11:131–167
10. Chawla N (2005) Data mining for imbalanced datasets: an overview. In: Maimon O, Rokach L (eds) The data mining and knowledge discovery handbook, pp 853–867, Springer, New York
11. Chawla N, Bowyer K, Hall L, Kegelmeyer W (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:341–378
12. Cost S, Salzberg S (1993) A weighted nearest neighbor algorithm for learning with symbolic features. Mach Learn J 10(1):1213–1228
13. Davis J, Goadrich M (2006) The relationship between Precision- Recall and ROC curves. In: Proceedings of the international conference on machine learning ICML, pp 233–240
14. Denil M, Trappenberg T (2011) A characterization of the combined effects of overlap and imbalance on the SVM classifier. In: Proceedings of CoRR conference, pp 1–10
15. Drummond C, Holte R (2006) Cost curves: an improved method for visualizing classifier performance. Mach Learn J 65(1):95–130
16. Elklan C (2001) The foundations of cost-sensitive learning. In: Proceedings of the international joint conference on artificial intelligence IJCAI-01, pp 63–66
17. Ester M, Kriegel H, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases. In: Proceedings of the international conference KDD'96, pp 226–231

18. Fernandez A, Garcia S, Herrera F (2011) Addressing the classification with imbalanced data: open problems and new challenges on class distribution. In: Proceedings of the HAIS conference (part. 1), pp 1–10
19. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2011) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern Part C: Appl Rev 99:1–22
20. Gamberger D, Boskovic R, Lavrac N, Groselj C (1999) Experiments with noise filtering in a medical domain. In: Proceedings of the 16th international conference on machine learning ICML'99, pp 143–151
21. Garcia S, Herrera F (2009) Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. Evol Comput 17(3):275–306
22. Garcia V, Sanchez JS, Mollineda RA (2007) An empirical study of the behaviour of classifiers on imbalanced and overlapped data sets. In: Proceedings of progress in pattern recognition, image analysis and applications 2007, LNCS, vol 4756, pp 397–406, Springer
23. Garcia V, Mollineda R, Sanchez JS (2008) On the k-nn performance in a challenging scenario of imbalance and overlapping. Pattern Anal Appl 11(3–4):269–280
24. Grzymala-Busse JW, Goodwin LK, Grzymala-Busse W, Zheng X (2000) An approach to imbalanced data sets based on changing rule strength. In: Proceeding of learning from imbalanced data sets, AAAI workshop at the 17th conference on AI, pp 69–74
25. Grzymala-Busse JW, Stefanowski J, Wilk S (2005) A comparison of two approaches to data mining from imbalanced data. J Intell Manufact 16(6):565–574
26. Gumkowski M (2014) Using cluster analysis to classification of imbalanced data. Master Thesis, Poznan University of Technology (supervised by Stefanowski J.)
27. Han H, Wang W, Mao B (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Proceedings of the ICIC, LNCS vol 3644, pp 878–887, Springer
28. Hand D (2009) Measuring classifier performance. A coherent alternative to the area under the ROC curve. Mach Learn J 42:203–231
29. He H, Garcia E (2009) Learning from imbalanced data. IEEE Trans Data Knowl Eng 21(9):1263–1284
30. He H, Ma Y (eds) (2013) Imbalanced learning. Foundations, algorithms and applications. IEEE—Wiley
31. Hido S, Kashima H (2009) Roughly balanced bagging for imbalance data. Stat Anal Data Min 2(5–6):412–426
32. Holte C, Acker LE, Porter BW (1989) Concept Learning and the problem of small disjuncts. In: Proceedings of the 11th IJCAI conference, pp 813–818
33. Japkowicz N (2001) Concept-learning in the presence of between-class and within-class imbalances. In: Proceedings of the Canadian conference on AI, pp 67–77
34. Japkowicz N (2003) Class imbalance: are we focusing on the right issue? In: Proceedings of the II workshop on learning from imbalanced data sets, ICML conference, pp 17–23:
35. Japkowicz N, Mohak S (2011) Evaluating learning algorithms: a classification perspective. Cambridge University Press, Cambridge
36. Japkowicz N, Stephen S (2002) Class imbalance problem: a systematic study. Intell Data Anal J 6(5):429–450
37. Jo T, Japkowicz N (2004) Class Imbalances versus small disjuncts. ACM SIGKDD Explor Newsl 6(1):40–49
38. Japkowicz N (2013) Assessment metrics for imbalanced learning. In: He H, Ma Y (eds) Imbalanced learning. foundations, algorithms and applications. IEEE—Wiley, pp 187–206
39. Kaluzny K (2009) Analysis of class decomposition in imbalanced data. Master Thesis (supervised by J. Stefanowski), Poznan University of Technology
40. Khoshgoftaar T, Van Hulse J, Napolitano A (2011) Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans Syst Man Cybern-Part A 41(3):552–568
41. Krawczyk B, Wozniak M, Schaefer G (2014) Cost-sensitive decision tree ensembles for effective imbalanced classification. Appl Soft Comput 14:544–562

42. Kubat M, Matwin S (1997) Addressing the curse of imbalanced training sets: one-side selection. In: Proceedings of the 14th international conference on machine learning ICML-97, pp 179–186
43. Kubat M, Holte R, Matwin S (1998) Machine learning for the detection of oil spills in radar images. Mach Learn J 30:195–215
44. Laurikkala J (2001) Improving identification of difficult small classes by balancing class distribution. Technical Report A-2001-2, University of Tampere
45. Lewis D, Catlett J (1994) Heterogeneous uncertainty sampling for supervised learning. In: Proceedings of 11th international conference on machine learning, pp 148–156
46. Lumijarvi J, Laurikkala J, Juhola M (2004) A comparison of different heterogeneous proximity functions and Euclidean distance. Stud Health Technol Inform 107(Part 2):1362–1366
47. Lopez V, Fernandez A, Garcia S, Palade V, Herrera F (2014) An Insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inform Sci 257:113–141
48. Maciejewski T, Stefanowski J (2011) Local neighbourhood extension of SMOTE for mining imbalanced data. In: Proceedings of the IEEE symposium on computational intelligence and data mining, pp 104–111
49. Maimon O, Rokach L (eds) (2005) The data mining and knowledge discovery handbook, Springer, New York
50. Maloof M (2003) Learning when data sets are imbalanced and when costs are unequal and unknown. In: Proceedings of the II workshop on learning from imbalanced data sets, ICML conference
51. Moore A, Pelleg D (2000) X-means: extending k-means with efficient estimation of the numbers of clusters. In: Proceedings of the 17th ICML, pp 727–734
52. Napierala K (2013) Improving rule classifiers for imbalanced data. Ph.D. Thesis. Poznan University of Technology
53. Napierala K, Stefanowski J (2012) The influence of minority class distribution on learning from imbalance data. In: Proceedings of the 7th conference HAIS 2012, LNAI vol 7209, pp 139–150, Springer
54. Napierala K, Stefanowski J (2012) BRACID: a comprehensive approach to learning rules from imbalanced data. J Intell Inform Syst 39(2):335–373
55. Napierala K, Stefanowski J, Wilk Sz (2010) Learning from imbalanced data in presence of noisy and borderline examples. In: Proceedings of 7th international conference RSCTC 2010, LNAI vol 6086, pp 158–167, Springer
56. Napierala K, Stefanowski J, Trzcielinska M (2014) Local characteristics of minority examples in pre-processing of imbalanced data. In: Proceedings of the ISMIS 2014, pp 123–132
57. Nickerson A, Japkowicz N, Milios E (2001) Using unsupervised learning to guide re-sampling in imbalanced data sets. In: Proceedings of the 8th international workshop on artificial intelligence and statistics, pp 261–265
58. Niemann U, Spiliopoulou M, Volzke, H, Kuhn JP (2014) Subpopulation discovery in epidemiological data with subspace clustering. Found Comput Decis Sci 39(4)
59. Prati R, Gustavo E, Batista G, Monard M (2004) Learning with class skews and small disjuncts. In: Proceedings of the SBIA 2004, LNAI vol 3171, pp 296–306, Springer
60. Prati R, Batista G, Monard M (2004) Class imbalance versus class overlapping: an analysis of a learning system behavior. In: Proceedings 3rd mexican international conference on artificial intelligence, pp 312–321
61. Parinaz S, Victor H, Matwin S (2014) Learning from imbalanced data using ensemble methods and cluster-based undersampling. In: Electronic Proceedings of the NFMCP 2014 workshop at ECML-PKDD 2014, Nancy
62. Saez JA, Luengo J, Stefanowski J, Herrera F (2015) Addressing the noisy and borderline examples problem in classification with imbalanced datasets via a class noise filtering method-based re-sampling technique. Inform Sci 291:184–203
63. Stefanowski J (2007) On combined classifiers, rule induction and rough sets. Trans Rough Sets 6:329–350

64. Stefanowski J (2013) Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. In: Ramanna S, Jain LC, Howlett RJ (eds) Emerging paradigms in machine learning, pp 277–306
65. Stefanowski J, Wilk Sz (2008) Selective pre-processing of imbalanced data for improving classification performance. In: Proceedings of the 10th international confernace DaWaK 2008. LNCS vol 5182, pp 283–292, Springer
66. Stefanowski J, Wilk Sz (2009) Extending rule-based classifiers to improve recognition of imbalanced classes. In: Ras ZW, Dardzinska A (eds) Advances in data management, Studies in computational intelligence, vol 223, pp 131–154, Springer
67. Ting K (1997) The problem of small disjuncts. Its remedy in decision trees. In: Proceedings of the 10th Canadian conference on AI, pp 91–97
68. Tomek I (1976) Two modifications of CNN. IEEE Trans Syst Man Commun 6:769–772
69. Van Hulse J, Khoshgoftarr T (2009) Knowledge discovery from imbalanced and noisy data. Data Knowl Eng 68:1513–1542
70. Van Hulse J, Khoshgoftarr T, Napolitano A (2007) Experimental perspectives on learning from imbalanced data. In: Proceedings of ICML, pp 935–942
71. Verbiest N, Ramentol E, Cornelis C, Herrera F (2012) Improving SMOTE with fuzzy rough prototype selection to detect noise in imbalanced classification data. In: Proceedings of the international conference IBERAMIA, pp 169–178
72. Weiss GM (2004) Mining with rarity: a unifying framework. ACM SIGKDD Explor Newsl 6(1):7–19
73. Weiss GM, Hirsh H (2000) A quantitative study of small disjuncts. In: Proceedings of the 17th national conference on artificial intelligence—AAAI00, pp 665–670
74. Weiss GM, Provost F (2003) Learning when training data are costly: the efect of class distribution on tree induction. J Artif Intell Res 19:315–354
75. Wilson DR, Martinez TR (1997) Improved heterogeneous distance functions. J Artif Intell Res 6:1–34
76. Zhu X, Wu X, Yang Y (2014) Error detection and impact-sensitive instance ranking in noisy data sets. In: Proceeding of the 19th national conference on AI, AAAI'04

Personal Privacy Protection in Time of Big Data

Marina Sokolova and Stan Matwin

Abstract Personal privacy protection increasingly becomes a story of privacy protection in electronic data format. Personal privacy protection also becomes a showcase of advantages and challenges of Big Data phenomenon. Accumulation of massive data volumes combined with development of intelligent Data Mining algorithms allows more data being analysed and linked. Unintended consequences of Big Data analytics include increasing risks of discovery new information about individuals. There are several approaches to protect privacy of individuals in the largeS. Matwin data sets, privacy-preserving Data Mining being an example. In this paper, we discuss *content-aware prevention* of data leaks. We concentrate on protection of personal health information (PHI), arguably the most vulnerable type of personal information. This paper discusses the applied methods and challenges which arise when we want to hold health information private. PHI leak prevention on the Web and on online social networks is our case study.

1 Introduction

Personal privacy protection increasingly becomes a story of privacy protection in electronic data. Personal privacy protection also becomes a showcase of advantages and challenges of Big Data phenomenon. Accumulation of massive data volumes of

M. Sokolova
School of Electrical Engineering and Computer Science, University of Ottawa,
Ottawa, Canada
e-mail: sokolova@uottawa.ca

M. Sokolova · S. Matwin
Institute for Big Data Analytics, Dalhousie University, Dalhousie, Canada

S. Matwin (✉)
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
e-mail: stan@cs.dal.ca

M. Sokolova
Faculty of Medicine, University of Ottawa, Ottawa, Canada

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_18

Table 1 PHI sources and data formats

Organizations			Individuals		
Main PHI sources	PHI data formats	Example	Main PHI sources	PHI data formats	Example
Electronic Health Record	Structured and semistructured	*Race:Caucas* *Gender:Fem* *Age: 3yr* *Diag_1:250.83* *Readmit: NO*	Online communities	Unstructured	*I have a family history of Alzheimer's disease. I have seen what it does and its sadness is a part of my life*

personal information (finances, health care, retail) combined with development of intelligent Data Mining algorithms allows personal data being analysed and linked in an innovative, but not always positive, way. Unintended consequences of Big Data analytics include increasing risks of discovering more information about individuals than was originally expected to be disclosed. For example, cross-site information aggregation can harvest information and link profiles on different social networking sites. By comparing the attributes from those profiles, individual profiles can be linked with high confidence. Furthermore, the attribute values can be used in the Web search to get other profiles of the individual [13].

As more data accumulates in various sources, large data breaches and inadvertent data leaks are becoming common: 2,164 incidents were reported in 2013,[1] a record number so far. Such incidents also expose a growing number of individual records, with 822 mln. records exposed in 2013. 48.5 % of all breaches had happened in US, 5.5 %—in UK, 2.7 %—in Canada, whereas other countries accounted for a smaller number of the incidents. 99 % of disclosed data was in electronic format. Further, we consider only data in the electronic format.

Responding to the public demand, data mining and machine learning research communities became involved in individual's privacy protection through Privacy-Preserving Data Mining, i.e., methods that perform data mining tasks in a way that strives to preserve privacy. The methods help to protect organizational and personal privacy in malicious breaches or inadvertent data leaks.

In this paper, we discuss content-aware prevention of inadvertent data leaks. We use protection of personal health information (PHI) as our on-going example. PHI refers to combination of personal information of an individual (e.g., date of birth) and information about one's health (e.g., diagnosis). PHI can be distributed by organizations and individuals alike; Table 1 provides data examples. All existing PHI protection methods are based on analysis of the data's content and as such they provide a comprehensive example for our study. Further in the paper, we provide more details on Data Leak Prevention (DLP) and PHI methods and techniques used to

[1] https://www.riskbasedsecurity.com/reports/2013-DataBreachQuickView.pdf.

safeguard data and information, and on challenges faced while solving the DLP and PHI protection problems. PHI leak detection on the Web and in online communities provides us with case studies of DLP applications, which has not yet been well studied. We discuss the problem challenges and suggest possible approaches to its solution. Discussion of future work and recommendations conclude the paper.

2 Principles of Data Security

Before we start a discussion of tools dedicated to content-aware DLP, we want to point out that basic good practices in personal data protection should start with systematic utilization of standard computer security techniques. These basic precautions apply to both data privacy breaches by malicious attackers, and to privacy violations by inadvertent disclosure (most often, human error). In particular, the following principles, if consistently reinforced across organizations, would be purposeful:

- compulsory encryption of all at rest data. This will protect against hacker attacks harvesting inactive data (e.g. the Target attack in 2013 and Home Depot attack in 2014 would likely be less significant if the data was encrypted; re-occurring attacks on E-commerce websites resulting in theft of credit card information would become inconsequential if data from past transactions were encrypted for storage).
- acknowledged tracing of private data access by authorized personnel. This would, to a large extent, avoid intentional privacy breaches internal to the organization. The approach has been used successfully in health care environments in major global centres to protect health-related data of recognized celebrities.
- education. In particular, sensitizing personnel with access to personal data about the dangers of placing this data in email, USB memory keys, laptops taken outside the organization and other Internet of Things devices.

We suggest that only after an organization is satisfied with ensuring the basic level of protection, described above and easily auditable, it should implement more targeted DLP techniques, particularly those meant to protect privacy in data made available for secondary use.

3 Data Leaks

There are many possible dimensions of data leaks. Leaks can be inadvertent or malicious, caused by insiders, third parties, or outsiders. Data can be in motion (e.g., traversing a network), at rest (e.g., stored on a computer for future use) or in use (e.g., processed on selected nodes of a network, but not traversing it at the moment). Most of our discussion involves data at rest. We will note when data in motion or in use are considered. In this study, we discuss prevention of inadvertent leaks. The leaks, unfortunately, had become expected as sharing the data among many insiders and outside service providers became the norm [7]. In 2013, more than 500 major data leaks were inadvertent. 66.4 % of inadvertent leaks have known causes; among those, information is disclosed on Web in 16.7 % and through email—in 14.7 %.

With 10 % of cost savings per patient from adoption of electronic medical/health records (EHR) [12], we observe the rising use of electronic data capture tools in clinical research. Large electronic repositories of personal health information (PHI) are being built up. For instance, in Ontario, Canada, the use of electronic medical records was doubled in 2006–2012[2]; in Europe, UK has the biggest EHR market, with $2.1bn in projected spending by the end of 2015.[3] We expect that the volume of PHI will grow with acceleration of home-delivered health care programs which are based on development of Internet of Things and the corresponding concept of Wear Your Own Device (WYOD) [26]. Heart monitoring implants and wearable sensors to detect falls are examples of biosensing devices used in home care.

PHI is in demand by academia and government agencies. The fast and wide-ranging acceptance of EHR guarantees its frequent secondary use, i.e. any use which differs from the original purpose of the data gathering. Data custodians (e.g., hospitals, doctor offices, government agencies) must assure its confidentiality and integrity and adhere to privacy and security rules, while making data available for secondary use. Integrating privacy protection in the EHR design can be a powerful tool [6], but might not be sufficient if used by its own.

Large medical data breaches and inadvertent disclosure are becoming common, substantially increasing concerns about data leaks. Survey of 43 US companies has shown that there are some distinct consequences of a data breach in health care.[4] In this industry more than others, customers notified of a data breach are more likely to discontinue association with companies that failed to secure sensitive data about them. While the average customer turnover due to a data breach was generally 3.6 %, customer turnover in financial services due to a data breach—5.5 %, in health care it was a much higher—6.5 %. And the cost of a health care breach, at $282 per record, was more than twice as high as that of the average retail breach at $131 per record (ibid.). The survey also showed that 44 % of data breaches occurred due to external causes involving third parties to whom the data has been transferred. PHI has leaked from health care providers, through documents sent by employees and medical students [11]. Data with insufficiently removed personal information can too cause PHI breaches [16].

Concerns about PHI leaks cause a moral retribution: patients may withhold information from a healthcare provider because of concerns over with whom the information might be shared, or how it might be used. The PHI breaches are potent legal issues, as PHI protection acts have been enabled by governments: Health Insurance

[2] https://www.infoway-inforoute.ca/index.php/news-media/2012-news-releases/use-of-electronic-medical-records-doubled-over-six-years.

[3] http://www.computerweekly.com/news/2240215175/UK-shows-biggest-take-up-of-electronic-Health-records-in-Europe.

[4] http://www.networkworld.com/news/2009/020209-data-breach.html.

Portability and Accountability Act, often known as HIPAA (US),[5] Personal Health Information Protection Act, or PHIPA (Ontario, Canada),[6] Data Protection Directive, or Directive 95/46/EC (EU).[7] It is thus imperative to have a global knowledge of best protocols and systems that protect privacy and confidentiality in data.

4 Personal Health Information: Definitions and Concepts

Personal Health Information (PHI) is a frequently used term usually defined through a set of meta-categories related to a person and his/her health conditions. The choice and number of meta-categories varies. For example, PHI is viewed as personally identifiable information (names, dates of birth and death, address, family relation) coupled with the health information about the person (diagnosis, prescriptions, medical procedure) [19]. In [10], PHI is identified through three meta-categories: exact personal identifiers (name, date of birth), quasi-identifiers (race, location) and sensitive health information (diagnosis, treatment). Studies by [29] have shown that references to health care providers (hospital, clinic) and professionals (doctors, nurses) frequently appeared in EHRs and revealed patient's health; thus such references should be considered as PHI. The Health Insurance Portability and Accountability Act (HIPAA) defines PHI in the most specific terms. It protects 18 PHI categories, e.g., geolocation, health plan beneficiary number. We list the categories in Table 2.

PHI can be stored and transferred in structured, semi-structured and unstructured text format. The former examples include pharmacy and admission records, the latter examples—free-text parts of patient's electronic medical records and letters, respectively. Health care organizations are main producers of PHI, and their networks and data bases are the major PHI depositories. PHI can be found on many online forums and social networks [8].

Data sharing and the secondary use of data are beneficial for patients, medical specialists, researchers in various domains and health care administration. At the same time, privacy laws mandate that individual's consent is obtained before the PHI data is shared. Obtaining the consent has been shown to lead to population selection bias [9]. To avoid pitfalls of both PHI disclosure and the consent request, data can go through a modification process, in which removal, alteration of generalization of personally identifiable information makes it difficult to identify a person from data. Such process is called anonymization. When patient records are anonymized, health care organizations are able to share such data without seeking preliminary patient's consent.

[5] http://www.hhs.gov/ocr/privacy/index.html.

[6] http://www.e-laws.gov.on.ca/html/statutes/english/elaws_statutes_04p03_e.htm.

[7] http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML.

Table 2 Health information protected by the Health Insurance Portability and Accountability Act

1. Names	10. Account numbers
2. All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code	11. Certificate/license numbers
3. Dates (other than year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89	12. Vehicle identifiers and serial numbers, including license plate numbers
4. Phone numbers	13. Device identifiers and serial numbers
5. Fax numbers	14. Web Uniform Resource Locators (URLs)
6. Electronic mail addresses	15. Internet Protocol (IP) address numbers
7. Social Security numbers	16. Biometric identifiers, including finger, retinal and voice prints
8. Medical record numbers	17. Full face photographic images and any com- parable images
9. Health plan beneficiary numbers	18. Any other unique identifying number, characteristic, or code

Three major types of data attributes are considered when anonymization is discussed [10]:

- explicit identifiers allow direct linking of an instance to a person (e.g., a cellular phone number or a drivers license number to its holder);
- quasi-identifiers, possibly combined with other attributes, may lead to other data sources and result in unique identification of a person; quasi-identifiers are often dates (birth, death, admission), locations (postal codes, hospital names, and regions), race, ethnicity, languages spoken, aboriginal status, and gender;
- non-identifying attributes are those for which there is no known inference linking to an explicit identifier.

Although all explicit identifiers are usually removed from the data as part of data preprocessing, this action may not be sufficient for privacy protection and other anonymization actions are required. For quasi-identifiers de-identification options are time-dependent, as algorithm's design is based on data sources and possible linkage of the day. At the same time, new attributes and data sources that can lead to a linkage to explicitly identifying attributes are constantly being engineered, thus, there will never be certainty about future de-identification as being shown in genomic data anonymization [14].

Commonly used Data Mining anonymization approaches may be categorized in the following groups:

Table 3 Comparison of k-anonymity, de-identification and differential privacy methods with respect to data format and data size

	Suitable data format			Data size required for performance		
	Structured	Unstructured	Semi-structured	Single record	Data set	Two data sets
k-anonymity	Yes	No	No	No	Yes	No
(text) De-identification	No	Yes	Yes	Yes	Yes	No
Differential privacy	Yes	No	No	No	Yes	Yes

1. k-anonymization. This approach stipulates that each record in a data set is similar to at least another k-1 records on the potentially identifying variables.
2. de-identification commonly refers to retrieval and extraction of PHI terms from semi-structured and unstructured texts.
3. differential privacy ensures that adding or removing a single dataset item does not substantially influence the outcome of data analysis.

The methods are evaluated through the information loss they incur and their efficiency. Applicability of those methods considerably varies. In Table 3 we exemplify data formats and data sizes necessary for each method.

5 Content-Aware Data Leak Prevention

Content-aware data leak prevention is the most technologically challenging part of data leak prevention. From the technology perspective, the methods discussed here belong to the field of Privacy-Preserving Data Mining; for survey refer to [16]. Content analysis of data is critical when two conflicting forces are at play: necessity to share the data and obligation to protect information the data holds. With the increase of secondary use of EHR (i.e., necessity to share) and the growing awareness of individual's privacy protection (i.e., obligation to protect), PHI leak prevention is a primary example of content-dependent leak prevention technology. Finances and security are other popular fields for deployment of content-aware information leak prevention tools. Such tools are deployed in banks, financial companies, government, i.e. organizations that keep and process sensitive and confidential information [28].

In a commonly used practice, DLP technologies monitor communications or networks to detect PHI leaks. When a leak is detected the affected individual or organization is notified, at which point they can take remedial action. DLP can either prevent a PHI leak or detect it after it happens. Some methods, applied to structured and unstructured data, remove explicit identifiers [27]. However, several well-popularized cases had shown that such removal may be insufficient for de-identification [16]. The majority of k-anonymity algorithms, which are applied to structured data sets, use generalization and suppression. Generalization reduces the value of attribute in precision or abstraction, at the same time reducing utility of the data. Suppression replaces a record’s value on an attribute with a missing value (this is called cell suppression), or in extreme cases a whole record is removed from the data set. Suppressing patient identifiable information may be insufficient when organizations disclose patient information, devoid of the sensitive information, to other agencies, e.g., DNA records transfer to publicly available data banks. In the case of distributed data, patients’ organization-visit patterns, or trails, can re-identify records to the identities from which they were derived.

While processing unstructured, free-form text data, tools apply Natural Language Processing, Information Extraction, and Machine Learning methods to enforce safer data management [17, 18]. Typically, a text de-identification method is either designed for one type of documents, e.g. discharge summaries, or a collection of documents coming from one health care facility. De-identification tasks apply Named Entity Recognition to identify, and possibly transform, personal information (names, age, location) [29]. Performance of algorithms is usually measured in numbers of individual term recognition. The reported metrics are F-score, Recall, and Precision [20]. Let us exemplify on the results of de-identification of nurse notes, i.e. short free-form texts, which contain abbreviations and professional terminology [21]. Stat De-id, an open-source de-identification tool, uses rules in search personal information. It can also use area-customised dictionaries (local person names, geographic and health care provider names). When the dictionaries are used, the overall Precision is 74.9 %. When the dictionaries are not used, the overall Precision decreases to 72.5 %. Performance substantially varies on identification of separate term categories. For person names, the use of the customized dictionaries is adverse: Precision = 73.1 % without the dictionaries and 72.5 %—with them. Locations, in contrast, considerably benefit from the use of the dictionaries: Precision increases from 84.0 to 92.2 % when the local information is available, Recall—from 37.0 to 97.0 %. Unfortunately, these results are considerably lower than manual de-identification, where the averaged Precision = 98.0 % (ibid.).

Development of new de-identification algorithms is constrained by limited access to training textual data. At that time, i2b2 data is the only set of clinical notes available for research.[8] Building such data is labor- and cost-consuming, as every sentence needs to be manually examined and anonymized. Use of advanced NLP methods can remedy the training data bottleneck as suggested in [5]. The researchers used distributional similarity for acquiring lexical information from notes typed by general practitioners. Lexical acquisition from ‘sensitive’ text enables effective processing of much larger datasets and, perhaps, will lead to construction of repository of training data.

Some techniques suggest that data holders securely collaborate through a third party: health care organizations keep all sensitive information in an encrypted state until the third party certifies that the data to be disclosed satisfies a formal data protection model [15]. It is important to note that those techniques are designed and

[8] https://www.i2b2.org/NLP/DataSets/Main.php.

applied to the data at rest. There is a general lack of tools which can detect and de-identify PHI in data sets in motion and in use. As with general DLP, different methods and tools may be required to safe-guard data in motion, at rest and in use. For the data in motion, if DLP is deployed to monitor email then a PHI alert can be generated before the email is sent. Early work in this area has been done by one of the authors [1, 3, 4]. This work investigated some of the issues related to privacy in an organizational setting, and proposed an approach based on information extraction and email content analysis in terms of compliance or non-compliance with organizational privacy rules. The approach emphasized a Role-based Access Control (RBAC) approach to checking compliance with privacy rules, as well as the need for ontologies which are needed to represent the roles and access privileges in an organization. A prototype has been developed in the university domain, where different agents (students, professors, administrative staff) have access to different information and can only communicate selected information to other agents. The prototype was based on a the real guidelines regarding Privacy Policy on the release of student information. For instance, student marks in a course C, accessible to administrative staff, can be mailed only to professors who teach C. Student's personal information cannot be emailed to other students or professors, etc.

For the data at rest, if DLP is used to monitor PHI leaks on the Internet (e.g., on peer-to-peer file sharing networks or on web sites), then the alerts pertain to leaks that have already occurred, at which point the affected individual or data custodian can attempt to contain the damage and stop further leaks. Other common shortcomings of the existing tools are:

- they apply to data sets built by health care providers; so far, few teams are actively involved in PHI leak prevention outside of health care providers' data sets.
- although patient records may provide information collected during multiple visits (this is called longitudinal data), there are currently techniques for the deidentification of numerical longitudinal medical and health records [24], but there are no longitudinal de-identification techniques for textual data; current text de-identification methods are designed for the cross-sectional data which is collected by observing patients at the same point of time or disregarding time differences.

6 PHI Leak Prevention on the Web

Emergence of the user-friendly Web 2.0 technologies imposed new challenges on DLP methods, due to increased volumes of data and an exponential increase in the number of data contributors. As online media becomes the main medium for the posting and exchange of information, analysis of this online data can contribute to studies of the general public's opinions on health-related matters. It has been shown that significant volumes of PHI can be found on the Web hubs (e.g., microblogs, online forums, social networks). Currently, 19–28 % of Internet users participate in online health discussions. Surveys of medical forum participants revealed that personal testimonials attract attention of up to 49 % of participants, whereas only 25 % of participants are motivated by scientific and practical content [2]. A 2011 survey of the US population estimated that 59 % of all adults have looked online for information about health topics such as a specific disease or treatment.[9]

Texts which host both PII and health information (PHI texts) are not analyzed regarding whether detected PHI is self-disclosed or published by a presumably trusted confidant. For example, a dedicated web site YourPrivacy separately considers PII in web posts, whereas health information is viewed within the scope of doctor-patient communication.[10] A few studies which analyzed information disseminated by health care professionals did not focus on PHI leaks, nor they analyzed large volumes of texts, thus do not have sufficient generalization power.

Indiscrimination between self-disclosure and confidentiality breaches means that PHI may remain on the Web, even when it has been detected due to the absence of rules prohibiting its self-disclosure. At the same time, the Web-posted information is freely available to 40 % of the world population, i.e. reaching 3 billion web users world-wide.[11] A concerned individual might not be aware about the PHI leak before the situation becomes harmful. For health care organizations, a timely detection of posted patient's PHI can prevent a breach of professional conduct. It can avert a possible negative impact for the company hosting the web site. Examples of host companies include such industry giants as Google, which hosts health discussion groups, and Yahoo!, which hosts health and wellness groups. Detection of the Web-posted PHI, thus, may prevent leaks in which legal and financial repercussions can be severe. Albeit hosting companies claim non-responsibility for displayed contents, courts may and do disagree.[12]

Identification of PHI leaks can be done by combining efforts of Natural Language Processing (NLP), Machine Learning (ML) and Software Engineering (SE), as it is often done with Web mining in general. Detection results must be delivered quickly. Efficiency is important, especially when health information is not innocuous (e.g., HIV, SARS, Ebola). We must make sensitive information unavailable as soon as possible. To that end, all relevant texts should be found and passed for manual processing. (People do not always agree on sensitivity assessment, so inter-judge agreement may be not all that high. We cannot definitely say that 100 % recall is indeed required.) Manual processing requires that false positives be as rare as possible, lest human control become lax with time.

There is another challenge: outside online medical communities, only a small percentage of text could have sensitive PHI information [25]. The learning algorithm, therefore, bases its decision on highly imbalanced input data, and the balance can change with time. Furthermore, possible inferences from publicly available data can have unintended consequences. To detect such consequences, security systems may use external knowledge repositories (e.g., Wikipedia, electronic dictionaries).

[9] http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx.

[10] http://www.yourprivacy.co.uk/.

[11] http://www.internetlivestats.com/internet-users/.

[12] http://en.wikipedia.org/wiki/District_of_Columbia_v.Heller.

Access to the additional information requires NLP- and ML-enhanced tools which are capable of extracting new knowledge from previously unseen data [8]. The last, but not the least, challenge is caused by stream-like data arrival. A continuous, steadfast current of information brings in unpredictability of content and context of individual data units. To detect PHI leak, a system has to process vast amount of un-predetermined information. To prevent adverse consequences, the detection should happen in the shortest possible time interval. Both requirements make the task of the Web-posted PHI leak prevention being computationally and methodologically complicated.

For a "smart" solution of seemingly intractable technological problem, we propose to concentrate on characteristics of PHI texts, instead. Each PHI file can be identified with a large, but restricted number of characteristics (e.g., names, addresses, diseases, drugs). We hypothesize that the PHI text detection can be compared to rare event detection. Those are extremely infrequent events whose characteristics make them or their consequences highly valuable. Such events appear with extreme scarcity and are hard to predict, although they are expected eventually to appear. The detection procedure can be based on finding distinguishable characteristics of the event. For detecting the Web-posted PHI, the proposed strategy is three-fold:

- develop, or adapt, a search algorithm which effectively seeks texts with a high potential of PHI (e.g., documents distributed by health organizations and health-care employees, lectures, presentations); the algorithm should effectively and efficiently discharge unrelated and irrelevant information;
- construct a set of the characteristics that can provide a high accuracy of PHI text detection; the sought after characteristics are syntactic and semantic indicators of PII and PHI in a text; for PII, we consider soft regular expressions for family relations, person's introductions, age-identifying events and dates, etc.; for PHI, we suggest the identification of term collocations in knowledge sources, used by health care organizations;
- apply named entity recognition through the use of specialized dictionaries; these dictionaries should contain information that is relevant for the given geographic area; note that geographic parameters are the most prominent pointers in person's identification.

7 PHL Leak Prevention in Online Communities

On the Web, most of PHI is disclosed through online communities. Despite the sensitive nature of PHI, participants do not always understand the risks of its online disclosure [30]. For example, a survey of participants from a popular forum Patients Like Me has found that 84 % of participants were comfortable with sharing their medical data with other participants, and 20 % with the whole world [31]. This is a potentially dangerous behaviour, as while participants may be in control of their messages, they cannot control the use of modern text analysis methodologies by third parties. By the means of Information Retrieval, Information Extraction, and Natural Language Processing, PHI can be harvested from the messages and further used to impinge with the privacy of the individuals. Three common tendencies illustrate this. First, online utilities facilitating the search of registered users: almost 50 % of disease-focused online communities had a tool performing such search. The most frequent user-searching queries were gender, age, username, geographic area, and disease [23]. Second, entice of disclosure of PHI: all medical forums encourage participants to share their experience of illness and seek advice from other participants. Such personal testimonials attract the attention of up to 49 % of participants [2]. Third, PHI-related marketing: 25 % of surveyed disease-focused networks were selling health-related products, including medical devices, parapharmaceutical products, and over-the-counter drugs and supplements [22]. Although online communities have developed and implemented policies which include the use of self-selected nicknames and editing of messages with unique identifiers (e.g., contact information, last names can be altered or removed), strong concerns have been expressed that such policies are not sufficient to protect the forum content, i.e., user messages. Additionally, many social media outlets gather and store participants' demographic and habitual information, records of past and future treatments, some outlets publish an aggregated statistics of participants [22].

To improve safety of this public space, we propose to develop evidence-based guidelines that help online users to avoid excessive PHI disclosure in online posts. One approach to alleviate some privacy-related problems is to protect PHI at the data origin or at the earliest opportunity before posting it. This approach is effective when the data origin venues are known and their numbers are limited, thus can be possibly tractable, e.g. messages posted on a given medical forum. However, it may not work in the distributed environment of online communities which enables users with multitude of platforms to post messages, either as a member of the group or as an individual. For example, from 41 social networks focusing on diseases, 14 networks claimed to be present as a group on at least one other social network (ibid.).

We propose to focus on those aspects of online privacy that are more directly at the user's control, i.e., privacy protection in user-written text. For instance, our guidelines will inform the users about potentially hazardous effects of disclosing personal identifiable information (e.g., full name, age), geographic personal pointers (e.g., home address, work location) and health information (e.g., diagnosis, health care unit). To achieve this goal, Social Mining technology can be used. It consists of the following steps: (i) identify online communities' characteristics that promote an excessive disclosure of PHI; (ii) determine the extent of problem, by investigating the proportion of users that excessively reveal PHI; and (iii) identify the demographics of the users that excessively reveal PHI. Necessity of such guidelines can be seen from an example of online text:

Message 1: [user's name] My brother *Jim Smith* …

Message 2: [user's name] "toss a few prayers up for my brother's roomie *John Doe* who in icu with a compound fracture of the knee".

Combined, both messages explicitly identify a person (*John Doe* who lives with *Jim Smith*) and his current health condition (a compound fracture of the knee treated in

ICU). The information makes the public aware that the said *John Doe*, being admitted to the hospital for a supposedly long time, neither currently lives at home nor is present at work. The person and health status link is dangerously excessive, e.g., the home can be targeted and vandalized during the absence of the residents. To prevent this excessive disclosure, our guidelines will inform users that "Personally identifiable information identifies a person among other individuals and usually consists of a person's given and family names and points to a location of the said person either through a home address or geographic relations (e.g., the name of the employer)". Two other guidelines could educate users on the steps to be taken to prevent the message from being over-informative:

- "Avoid linking the name of the person and his/her diagnosis with the health care organization". Aware of this guideline, users would instead write "toss a few prayers up for my brother's roomie who has a compound fracture of the knee".
- "Avoid linking the name of the person and his/her physical geographic location". Aware of this guideline, users would instead write "toss a few prayers up for *John Doe* who has a compound fracture of the knee".

By being informed what constitutes personally identifiable information, users will be aware of potential dangers of posting given and family names and geographic pointer of a person. After being guided on how to avoid an inadvertent disclosure of private information and personally identifiable information, users will not jeopardize their safety and will improve security of their adobes. The users will also be sensitized to what information they can post about others, thus, reducing the chances of inadvertent confidentiality breaches.

8 Final Remarks and Future Work

This paper summarizes some of the existing challenges in privacy protection within realms of Big Data, when personal information of individuals can be found, cross-examined and aggregated from many different data sources. We have discussed merits of content-aware prevention of *inadvertent* data leaks and its application to protection of personal health information. We have illustrated that successful implementation of leak prevention methods relies on solution of several Data Mining problems. In this paper, we have proposed pro-active content-aware prevention of personal health information leaks which can work well in online communities.

So far, the early attempts at solutions are developed. If robust leak prevention methods are to be built in order to avoid harmful privacy violation incidents, there is a strong need to work on a number of important issues. We outline some of the issues below, and welcome a discussion on others.

- One of such issues is the detection of personally identifying information and health information in multi-modal data, e.g., data containing text and images, web page data combining text, images and web site links, and potential inference from external web site links, e.g., Wikipedia.

- Another issue concerns data anonymization specifically in the healthcare context. What original data properties should be preserved, guaranteeing the quality and utility of data for secondary data use (e.g. research)? What quasiidentifiers should be changed and how (for example, postal code can be changed in a different way depending on the country). What are the appropriate techniques, especially for longitudinal data? Techniques such as random name and address substitution, use of non-existing names and addresses are popular now, but were they properly assessed? What are the criteria and measures of successful anonymization, leading to measures of degree of privacy protection?
- This leads us to system development, testing, and deployment issues. They include the challenges of combining Software Engineering, Machine Learning and Natural Language Processing components and obtaining efficient, scalable systems. There are all-important issues of system testing and evaluation of systems performance and user-acceptance: training and testing, failure detection, applicability of performance measures, e.g., efficiency, effectiveness, robustness, accuracy, precision, and generalization of results.
- Finally, there are important non-technical issues that need to be addressed for any proposed technical solution. They include such questions as where does responsibility shift from data custodian to the PHI DLP system creators? What are responsibilities of third parties, e.g., secondary data users? How to train and sensitize users of the DLP systems?

We hypothesize that solution of the following problems is important for future development of Privacy-preserving Data Mining methods:

- Balancing individual privacy and collective privacy within shared data; for example, in the same hospital data, privacy concerns can be different for patients diagnosed with “good” disease and those who are diagnosed with “bad” disease.
- Estimating feasibility of personalized privacy; we can start with looking into two questions: personal pros and cons can and will change with time—how to deal with old data sets? can existing techniques generalize well from personalized records?
- Search for measures and metrics assessing algorithm’s performance in privacy protection.

As for immediate PPDM tasks, we suggest to build repository of data sets, similar to UCI and develop a tool kit similar to WEKA. We believe that these are socially and economically important research areas, worthy of attention of the privacy research community.

References

1. Armour Q, Elazmeh W, Nour El-Kadri N, Japkowicz N, Matwin S (2005) Privacy compliance enforcement in Email. Adv Artif Intell 18:194–204 (Springer)

2. Balicco L, Paganelli C (2011) Access to health information: going from professional to public practices. In: 4th International conference on information systems and economic intelligence, p 135
3. Boufaden N, Elazmeh W, Ma Y, Matwin S, El-Kadri N, Japkowicz N (2005) PEEP- an information extraction base approach for privacy protection in Email. CEAS
4. Boufaden N, Elazmeh W, Matwin S, Japkowicz N (2005) PEEP- privacy enforcement in Email project. In: Third annual conference on privacy, security and trust, pp 257–260
5. Carroll J, Koeling R, Puri S (2012) Lexical acquisition for clinical text mining using distributional similarity. In: Computational linguistics and intelligent text processing. Springer, New York, pp 232–246
6. Cavoukian A, Alvarez A (2012) Embedding privacy into the design of EHRs to enable multiple functionalities—Win/Win. Canada Health Infoway
7. Davenport T, McNeill D (2014) Analytics in healthcare and the life sciences. International Institute for Analytics
8. Ghazinour K, Sokolova M, Matwin S (2013) Detecting health-related privacy leaks in social networks using text mining tools. Adv Artif Intell 26:25–39 (Springer)
9. Harris A, Teschke K (2008) Personal privacy and public health: potential impacts of privacy legislation on health research in Canada. Can J Public Health 99:293–296
10. Jafer Y, Matwin S, Sokolova M (2014) Task oriented privacy preserving data publishing using feature selection. Adv Artif Intell 27:143–154 (Springer)
11. Johnson E (2009) Data hemorrhages in the health-care sector. In: Financial cryptography and data security, Springer, pp 71–89
12. Kazley A, Simpson A, Simpson K, Teufel R (2014) Association of electronic health records with cost savings in a national sample. Am J Manag Care 183–190
13. Li F, Zou X, Liu P, Chan J (2011) New threats to health data privacy. BMC Bioinf. doi:10.1186/1471-2105-12-S12-S7
14. Malin B (2005) An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. J Am Med Inform Assoc 12:28–34
15. Malin B (2010) Secure construction of k-unlinkable patient records from distributed providers. Artif Intell Med 48:29–41
16. Matwin S, Szapiro T (2010) Data privacy: from technology to economics. In: J Koronacki et al (eds) Advances in machine learning II. Springer, New York, pp 43–74
17. McCoy A, Wright A, Eysenbach G, Malin B, Patterson E, Xu H et al (2013) State of the art in clinical informatics: evidence and examples. In: IMIA Yearbook of Medical Informatics, pp 1–8
18. Meystre S, Friedlin F, South B, Shen S, Samore M (2010) Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. doi:10.1186/1471-2288-10-70
19. Mitiku T, Tu K (2008) ICES report: using data from electronic medical records: theory versus practice. Healthc Q 11(2):23–25
20. Muqun L, Carrell D, Aberdeen J, Hirschman L, Malin B (2014) De-identification of clinical narratives through writing complexity measures. Int J Med Inform 83(10):750–767
21. Neamatullah I, Douglass M, Lehman L, Reisner A, Villarroel M, Long W et al (2008) Automated de-identification of free-text medical records. Med Inform Decis Mak 8(32):24–32
22. Orizio G, Schulz P, Gasparotti C, Caimi L (2010) The world of e-patients: a content analysis of online social networks focusing on diseases. Telemed J E Health 16(10):1060–1066
23. Richter J, Becker A, Schalis H, Koch T, Willers R, Specker C et al (2011) An ask-the-expert service on a rheumatology web site: who are the users and what did they look for? Arthritis Care Res 63(4):604–611
24. Sehatkar M (2014) Towards a privacy preserving framework for publishing longitudinal data (Ph.D. thesis). University of Ottawa
25. Sokolova M, El Emam K, Arbuckle L, Neri E, Rose S, Jonker E (2012) P2P Watch: personal health information detection in peer-to-peer file sharing networks. J Med Internet Res. http://dx.doi.org/10.2196/jmir.1898

26. Swan M (2012) Sensor Mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. J Sens Actuator Netw 1(3):217–253
27. Sweeney L (2006) Protecting job seekers from identity theft. IEEE Internet Comput 10(2):74–78
28. Tahboub R, Saleh Y (2014) Data leakage/loss prevention systems. Comput Appl Inform Syst 1–6 (IEEE)
29. Uzuner O, Luo Y, Szolovits P (2007) Evaluating the state-of-the-art in automatic de-indentification. J Am Med Inform Assoc 14(5):550–563
30. Van der Velden M, El Emam K (2012) Not all my friends need to know: a qualitative study of teenage patients, privacy and social media. J Am Med Inform Assoc 20(1):16–24
31. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T et al (2010) Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res. http://dx.doi.org/10.2196/jmir.1549

Data Based Modeling

James R. Thompson

Abstract Statisticians spend a great deal of time coming up with tests that are frequently useless in practice and then proving the asymptotic optimality of the tests under the assumption of conditions that do not actually exist. There is another approach: we can use the data to build models. This is the goal of Tukey's "Exploratory Data Analysis." In this paper I will be examining alternatives to the Neoclassical Analysis of the stock market, the dominant view still in schools of business, data notwithstanding. Then I will give a brief analysis of the breakdown of America's public health service in stopping the progression of the AIDS epidemic and demonstrate that simply closing down the gay bathhouses would have prevented AIDS from progressing from an endemic to a full blown epidemic which has already killed more Americans than died in the two world wars.

Keywords Capital market line · Simugram · Maxmedian · AIDS

1 The Introduction

There is a tendency for mathematicians and statisticians (including "applied" ones) to believe that in the vast literature of theoretical tools, there must be one appropriate to the problem at hand. This is generally not the case. This fact has been emphasized by Marian Rejewski, who cracked the Enigma code used by the German armed forces, and most sophisticatedly by the German Navy. Dr. Rejewski was not just a theoretical mathematician, but one who had four years of statistical training at Gottingen. Given the task in 1931, he tried the rich panolpy of techniques he had learnt to no effect. Bydgoszcz, where he attended high school was part of the German chunk of partitioned Poland. So as a cadet in high school he learned much about the eccentricities used in military and naval German reports. For example, memos started with a line beginning with "Von" followed by a second line starting with "Zu". Then

J.R. Thompson (✉)
Department of Statistics, Rice University, Houston, TX 77251-1892, USA
e-mail: thomp@rice.edu

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_19

turning linguist and cultural sociologist, Rejewski built up a template of forms that must be used in military discourse. At the end of the day, he had reduced the number of feasible combinations in Enigma from 10^{92} to a manageable 100,000. Every time the code was changed by the Germans, using a few dozen cypher clerks, the Rejewski team could come up with the settings used in the new format in a week or so. (It should be noted in passing that the submarine codes could only be changed when submarines were docking at German occupied ports and that the SS never departed from the original settings of 1932).

The British have always minimized the fact that it was the Poles who cracked Enigma. However, Rejewski and his crew saved the British from a starvation-induced peace with the Nazis. Rejewski's filtering "bombe" was the first digital computer and the coding is correctly viewed as proto-Unix. It is usually the case that real world problems require stepping outside the standard tool boxes of mathematics and statistics.

2 If Only the Market Were a Martingale (But It Is Not)

One way to express the Weak Form of the Efficient Market Hypothesis is to require that stocks have martingale structure, i.e., for a stock $S(t)$, the expected value at any future time $t + r$ is $S(t)$. In other words, a stock which has been going up for the last 10 sessions is no more worthy an investment than a stock which has gone down for the last 10 sessions. This is counterintuitive, but has been the basis of several Nobel Prizes in Economics. One of these belongs to William Sharpe for his Capital Market Theory [1, 2].

If we may assume that investors behave in a manner consistent with the Efficient Market Hypothesis, then certain statements may be made about the nature of capital markets as a whole. Before a complete statement of capital market theory may be advanced, however, certain additional assumptions must be presented:

1. The μ and σ of a portfolio adequately describe it for the purpose of investor decision making $[U = f(\sigma, \mu)]$.
2. Investors can borrow and lend as much as they want at the riskless rate of interest.
3. All investors have the same expectations regarding the future, the same portfolios available to them, and the same time horizon.
4. Taxes, transactions costs, inflation, and changes in interest rates may be ignored.

Under the above assumptions, all investors will have identical opportunity sets, borrowing and lending rates $(r_L = r_B)$ and, thus, identical optimal borrowing-lending portfolios, say X (see Fig. 1).

The borrowing-lending line for the market as whole is called the *Capital Market Line*. The "market portfolio" (X) employed is the total universe of available securities weighted by their total stock value relative to all the stocks in the market universe (called the *market portfolio*) by the reasoning given above. The CML is linear and it represents the combination of a risky portfolio (X) and a riskless security (a T$-$Bill).

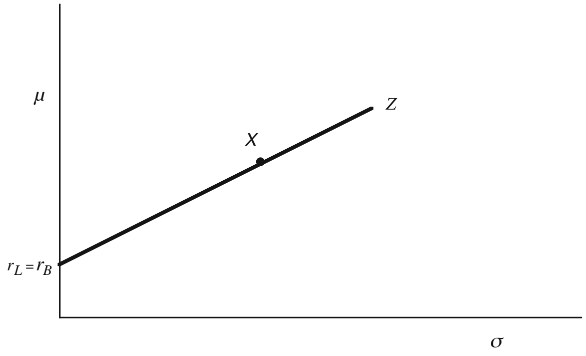


Fig. 1 The capital market line

One use made of the CML is that its slope provides the so-called *market price of risk*, or, that amount of increased return required by market conditions to justify the acceptance of an increment to risk, that is

$$\text{slope} = \frac{\mu(X) - r}{\sigma(X)}.$$

The simple difference $\mu(X) - r$ is called the *equity premium*, or the expected return differential for investing in risky equities rather than riskless debt.

This very elegant result of Sharpe indicates that one simply cannot do better than invest along the Sharpe Superefficient Frontier (CML). Of course, if one wishes to invest on "autopilot" there are ways to do so. John Bogle has effectively and non-mathematically argued [3] that the value of investment counsellors is, in general, not worth their fees. Many years ago, he founded the Vanguard S&P 500 fund (among others) which maintains a portfolio balanced according to the market cap values of each of the members of the Standard and Poor selected basket of top 500 stocks. Thus the weight of investment in the $i'th$ stock would be

$$w_i = \frac{V_i}{\Sigma V_j} \tag{1}$$

where V_i is the total market value of all the stocks in company i. Interestingly, Bogle's strategy is actually very close to the "total market index fund" suggested by Nobel laureate William Sharpe. However, Thompson et al. [4] took a backlook at 50,000 randomly selected portfolios from the 1,000 largest market cap stocks over a period of 40 years. They discovered that over over half lie above the CML. How it has been that EMH enthusiasts apparently failed to crunch the numbers is a matter of conjecture. Nor is this result surprising, since the Standard and Poor Index fund over this period has averaged an annual return of somewhat in excess of 10 % while Buffett's Berkshire-Hathaway has delivered over 20 % (Fig. 2).

When we see that randomnly selected portfolios frequently lie above the Capital Market Line, we are tempted to see what happens when we make a selection based on equal weights of, for example, the Standard and Poor 100. We shall also

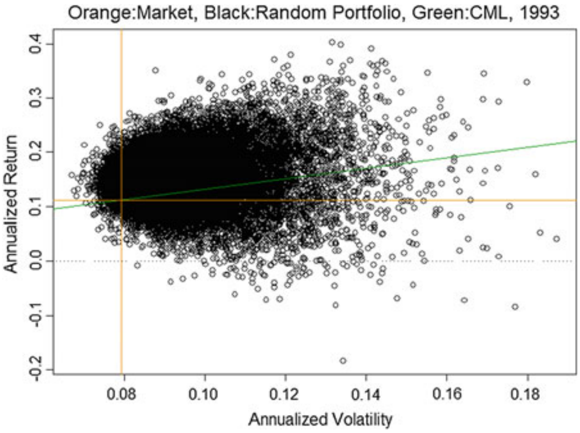


Fig. 2 Randomly selected portfolios beating the super efficient frontier portfolios

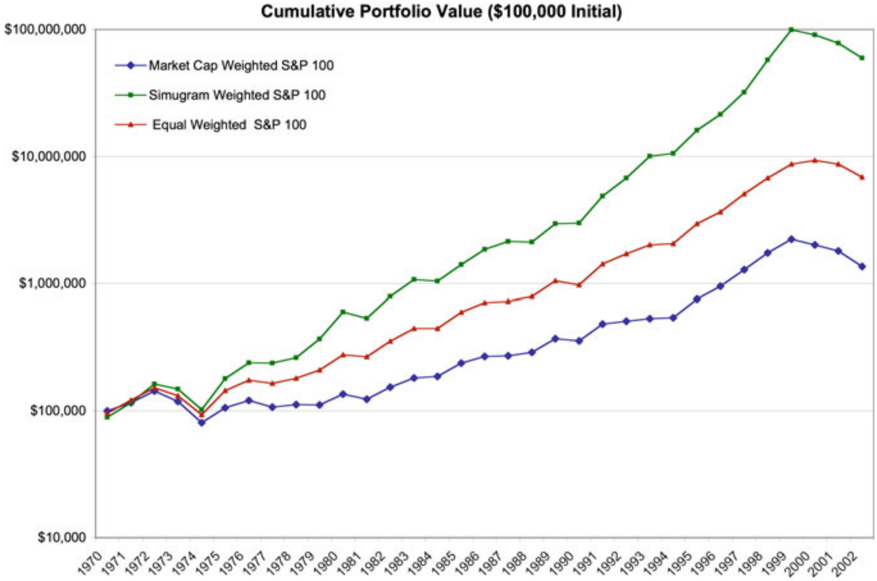


Fig. 3 Market cap weight versus equal weight

demonstrate the results of Thompson's patented Simugram portfolio selection algorithm [5]. Space does not permit a discussion of this algorithm. Suffice it to say that though quite different from the fundamental analysis of Buffet, it achieves roughly the same results. During the economic shocks caused by the market collapse of 2008–2009, both the Simugram and the analysis of Buffett proved themselves nonrobust against massive intervention of the Federal Reserve Bank to save the large banks. (Now that QE3 has ended, Simugram appears to be working again).

We show such a comparison in Fig. 3.

3 “Everyman’s” MaxMedian Rule for Portfolio Management

If index funds, such as Vanguard’s S&P 500 are popular (and with some justification they are), this is partly due to the fact that over several decades the market cap weighted portfolio of stocks in the S&P 500 of John Bogle (which is slightly different that a total market fund) has small operataing fees, currently, less than 0.1 % compared to fund management rates typically around 40 times that of Vanguard. And, with dividends thrown in, it produces around a 10 % return. Many people prefer large cap index funds like those of Vanguard and Fidelity. The results of managed funds have not been encouraging overall, although those dealing with people like Peter Lynch and Warren Buffet have done generally well. John Bogle probably did not build his Vanguard funds because of any great faith in fatwahs coming down from the EMH professors at the University of Chicago. Rather, he was arguing that investors were paying too much for the “wisdom” of the fund managers. There is little question that John Bogle has benefited greatly the middle class investor community.

That being said, we have shown earlier that market cap weighted funds do no better (actually worse) than those selected by random choice. It might, then, be argued that there are nonrandom strategies which the individual investor could use to his/her advantage. For example if one had invested in the stocks *with equal weight* in the S&P 100 over the last 40 years rather than by weighting according to market cap, he would have experienced a significantly higher annual growth (our backtest revealed as much as a 5 % per year difference in favor of the equal weighted portolio). We remind the reader that the S&P 100 universe has been selected by fundamental analysis from the S&P 500 according to fundamental analysis and balance. Moreover, the downside losses in bad years would have been less than with a market cap weighted fund. It would be nice if we could come up with a strategy which kept only 20 stocks in the portfolio. If one is into managing ones own portfolio, it would appear that Baggett and Thompson [6] did about as well with their MaxMedian Rule as the equal weight of the S&P 100 using a portfolio size of only 20 stocks. I am harking back to the old morality play of “Everyman” where the poor average citizen moving through life is largely abandoned by friends and advisors except for *Knowledge* who assures him “Everyman, I will accompany you and be your guide.”

The MaxMedian Rule [6] of Baggett and Thompson, given below, is easy to use and appears to beat the Index, on the average, by up to an annual multiplier of 1.05, an amount which is additionally enhanced by the power of compound interest. Note that $(1.15/1.10)^{45} = 7.4$, a handy bonus to one who retires after 45 years. A purpose of the MaxMedian Rule was to provide individual investors with a tool which they could use and modify without the necessity of massive computing. Students in my classes have developed their own paradigms, such as the MaxMean Rule. In order to use such rules, one need only purchase for a very modest one time fee the Yahoo base *hquotes* program from hquotes.com. (The author owns no portion of the hquotes company).

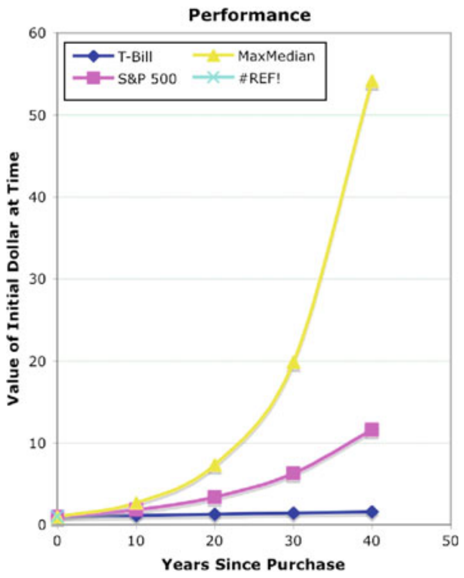


Fig. 4 A comparison of three investment strategies

The MaxMedian Rule

1. For the 500 stocks in the S&P 500 look back at the daily returns $S(j,t)$ for the preceding year
2. Compute the day to day ratios $r(j,t) = S(j,t)/S(j,t-1)$
3. Sort these for the year's trading days
4. Discard all r values equal to one
5. Look in the 500 medians of the ratios
6. Invest equally in the 20 stocks with the largest medians
7. Hold for one year, then liquidate.

In Fig. 4 we examine the results of putting one present value dollar into play in three different investments: 5 % yielding T-Bill, S&P 500 Index Fund, MaxMedian Rule. First, we shall do the investment simply without avoiding the intermediate taxing structure. The assumptions are that interest income is taxed at 35 %; capital gains and dividends are taxed at 15 %; and inflation is 2 %. As we see, the T-Bill invested dollar is barely holding its one dollar value over time. The consequences of such an investment strategy are disastrous as a vehicle for retirement. On the other hand, after 40 years, the S&P 500 Index Fund dollar has grown to 11 present value dollars. The MaxMedian Rule dollar has grown to 55 present value dollars. Our investigations indicate that the MaxMedian Rule performs about as well as an equal weighted S&P 100 portfolio, though the latter has somewhat less downside in bad years. Of course, it is difficult for the individual investor to buy into a no load equal weight S&P 100 index fund. So far as the author knows, none currently exist, though equal weighted S&P 500 index funds do (the management fees seem to be in the 0.50 % range). For reasons not yet clear to the author, the advantage of the equal weight S&P index fund is only 2 % greater than that of the market cap weight S&P 500. Even so, when one looks at the compounded advantage over 40 years, it appears to be roughly a factor of two. It is interesting to note that the bogus Ponzi scheme of Maidoff claimed returns which appear to be legally attainable either by the MaxMedian Rule or the equal weight S&P 100 rule. This leads the author to the conclusion that most of the moguls of finance and the Federal Reserve Bank have very limited data analytical skills or even motivation to look at market data.

3.1 Investing in a 401-k

Money invested in a 401-k plan avoids all taxes until the money is withdrawn, at which time it is taxed at the current level of tax on ordinary income. In Table 1, we demonstrate the results of adding an annual inflation adjusted $5,000 addition to a 401k for 40 years, using different assumptions of annual inflation rates. ($5,000 is very modest but that sum can be easily adjusted.) All values are in current value dollars.

We recall that when these dollars are withdrawn, taxes must be paid. So, in computing the annual cost of living, one should figure in the tax burden. Let us suppose the cost of living including taxes for a family of two is $70,000 beyond Social Security retirement checks. We realize that the 401-k portion which has not been withdrawn will continue to grow (though the additions from salary will have ceased upon retirement). Even for the unrealistically low inflation rate of 2 % the situation is not encouraging for the investor in T-bills. Both the S&P Index holder and the Max Median holder will be in reasonable shape. For the inflation rate of 5 %, the T-bill holder is in real trouble. The situation for the Index Fund holder is also risky. The holder in the MaxMedian Rule portfolio appears to be in reasonable shape. Now, by historical standards, 5 % inflation is high for the USA. On the other hand, we observe that the decline of the dollar against the Euro during the Bush Administration was as high as 8 % per year.

Hence, realistically, 8 % could be a possibility to the inflation rate for the future in the United States. In such a case, of the four strategies considered, only the return available from the MaxMedian Rule leaves the family in reasonable shape. Currently, even the Euro is inflation-stressed due to the social welfare excesses of some of the

Table 1 40 year end results of three 401-k strategies

Inflation	2 %	3 %	5 %	8 %
T-Bill	447,229	292,238	190,552	110,197
S&P Index	1,228,978	924,158	560,356	254,777
MaxMedian	4,660,901	3,385,738	1,806,669	735,977

Eurozone members. From a societal standpoint, it is not necessary that an individual investor achieve spectacular returns. What is required is effectiveness, robustness, transparency, and simplicity of use so that the returns will be commensurate with the normal goals of families: education of children, comfortable retirement, etc. Furthermore, it is within the power of the federal government to bring the economy to such a pass where even the prudent cannot make do. The history of modern societies shows that high rates of inflation cannot be sustained without some sort of revolution, such as that which occurred at the end of the Weimar Republic. Unscrupulous bankers encourage indebtedness on the unwary, taking their profits at the front end and leaving society as a whole to pick up the bill. Naturally, as a scientist, I would hope that the empirical rules such as the MaxMedian approach of Baggett and Thompson will lead to fundamental insights about the market and the economy more generally. Caveat: The MaxMedian Rule is freeware not quality assured or extensively tested. If you use it, remember what you paid for it. The goal of the MaxMedian Rule is to enable the individual investor to develop his or her own portfolios without the assistance of generally overpriced and underachieving investment fund managers. The investor gets to use all sorts of readily available information in public libraries, e.g., *Investors Business Daily*. Indeed, many private investors will subscribe to *IBD* as well as to other periodicals. Obviously, even if a stock is recommended by the MaxMedian rule (or any rule) and there is valuable knowledge, such as that the company represented by the stock is under significant legal attack for patent infringement, oil spills, etc., exclusion of the stock from the portfolio might be indicated. The bargain brokerage *Fidelity* provides abundant free information for its clients and generally charges less than 8 dollars per trade.

Obviously, one might choose rather a MaxMean rule or a Max 60 Percentile rule or an equal weight Index rule. The MaxMedian was selected to minimize the optimism caused by the long right hand tails of the log normal curves of stock progression. MaxMean is therefore more risky. There are many which might be tested by a forty year backtest. My goal is not to push the MaxMedian Rule or the MaxMean Rule or the equal weight S&P 100 rule or any rule, but rather allow the intelligent investor to invest without paying vast sums to overpriced and frequently clueless MBAs. If, at the end of the day, the investor chooses to invest in market cap based index funds, that is suboptimal but not ridiculous. What is ridiculous is not to work hard to understand as much as practicable about investment. This chapter is a very good start. It has to be observed that at this time in history, investment in US Treasury Bills or bank cds would appear to be close to suicidal. Both the Federal Reserve and the investment banks are doing the American middle class no good service. 0.1 % return on Treasury Bills is akin to theft, and what some of the investment banks do is akin to robbery. By lowering the interest rate to nearly zero, the Federal Reserve has damaged the savings of the average citizen and laid the groundwork for future high inflation. The prudent investor is wise to invest in stocks rather than in bonds.

I have no magic riskless formula for getting rich. Rather, I shall offer some opinions about alternatives to things such as buying T-Bills. Investing in market cap index funds is certainly suboptimal. However, it is robustness and transparency rather than optimality which should be the goal of the prudent investor. It should be remembered that most investment funds do charge the investor a fair amount of his/her basic investment whatever be the results. The EMH is untrue and does not justify investment in a market cap weighted index fund. However, the fact is that, with the exception of such gurus as Warren Buffett and Peter Lynch, the wisdom of the professional market forecaster seldom justifies the premium of the guru's charge. There are very special momentum based programs (on one of which the author holds a patent), in which the investor might do well. However, if one simply manages one's own account, using MaxMean or MaxMean within an IRA, it would seem to be better than trusting in gurus who have failed again and again. Berksire-Hathaway has proved to be over the years a vehicle which produces better than 20 % return. For any strategy that the investor is considering, backtesting for, say, 40 years, is a very good idea. That is not easy to achieve with equal weight funds, since they have not been around very long. Baggett and Thompson had to go back using raw S&P 100 data to assess the potential of an S&P 100 equal weight fund. If Bernie Maidoff had set up such a fund, he might well have been able to give his investors the 15 % return he promised but did not deliver.

The United States government has been forcing commercial banks to grant mortgage loans to persons unlikely to be able to repay them, and its willingness to allow commercial banks to engage in speculative derivative sales, is the driving force behind the market collapse of the late Bush Administration and the Obama Administration. Just the war cost part of the current crisis due to what Nobel Laureate Joseph Stiglitz has described as something beyond a three trillion dollar war in the Middle East has damaged both Berkshire-Hathaway's and other investment strategies. To survive in the current market situation, one must be agile indeed. Stiglitz keeps upping his estimates of the cost of America's war in the Middle East. Anecdotally, I have seen estimates as high as six trillion dollars. If we realize that the cost of running the entire US Federal government is around three trillion dollars per year, then we can see what a large effect Bush's war of choice has had on our country's aggregate debt. This fact alone would indicate that a future damqging inflation is all but certain. To some extent, investing in the stock market could be viewed as a hedge against inflation.

In the next section, we will examine another cause of denigration and instability in the economy, the failure of the Centers for Disease Control to prevent the AIDS endemic from becoming an AIDS epidemic.

4 AIDS: A New Epidemic for America

In 1983, I was investigating the common practice of using stochastic models in dealing with various aspects of diseases. Rather than considering a branching process model for the progression of a contagious disease, it is better to use differential equation models of the mean trace of susceptibles and infectives. At this time the disease had infected only a few hundred in the United States and was still sometimes referred to as GRIDS (Gay Related Immunodeficiency Syndrome). The more politically correct name of AIDS soon replaced it.

Even at the very early stage of an observed United States AIDS epidemic, several matters appeared clear to me:

- The disease favored the homosexual male community and outbreaks seemed most noticeable in areas with sociologically identifiable gay communities.
- The disease was also killing (generally rather quickly) people with acute hemophilia.
- Given the virologist's maxim that there are no new diseases, AIDS in the United States had been identified starting around 1980 because of some sociological change. A disease endemic under earlier norms, it had blossomed into an epidemic due to a change in society.

At the time, which was before the HIV virus had been isolated and identified, there was a great deal of commentary both in the popular press and in the medical literature (including that of the Centers for Disease Control) to the effect that AIDS was a new disease. Those statements were not only false but were also potentially harmful. First of all, from a practical virological standpoint, a new disease might have as a practical implication genetic engineering by a hostile foreign power. This was a time of high tension in the Cold War, and such an allegation had the potential for causing serious ramifications at the level of national defense.

Secondly, treating an unknown disease as a new disease essentially removes the possibility of stopping the epidemic sociologically by simply seeking out and removing (or lessening) the cause(s) that resulted in the endemic being driven over the epidemiological threshold.

For example, if somehow a disease (say, the Lunar Pox) has been introduced from the moon via the bringin in of moon rocks by American astronauts, that is an entirely different matter than, say, a mysterious outbreak of dysentery in St. Louis. For dysentery in St. Louis, we check food and water supplies, and quickly look for "the usual suspects"—unrefrigerated meat, leakage of toxins into the water supply, and so on. Given proper resources, eliminating the epidemic should be straightforward.

For the Lunar Pox, there are no usual suspects. We cannot, by reverting to some sociological *status quo ante*, solve our problem. We can only look for a bacterium or virus and try for a cure or vaccine. The age-old way of eliminating an epidemic by sociological means is difficult—perhaps impossible.

In 1982, it was already clear that the United States public health establishment was essentially treating AIDS as though it were the Lunar Pox. The epidemic was at levels hardly worthy of the name in Western Europe, but it was growing. Each of the European countries was following classical sociological protocols for dealing with a venereal disease. These all involved some measure of defacilitating contacts between infectives and susceptibles. The French demanded bright lighting in gay "make-out" areas. Periodic arrests of transvestite prostitutes in the Bois de Boulogne were widely publicized. The Swedes took much more draconian steps, mild in comparison with those of the Cubans. The Americans took no significant sociological steps at all.

However, as though following the Lunar Pox strategy, the Americans outdid the rest of the world in money thrown at research related to AIDS. Some of this was spent on isolating the unknown virus. However, it was the French, spending pennies to the Americans' dollars, at the Pasteur Institute who first isolated HIV. In the intervening 30 years since isolation of the virus, no effective vaccine or cure has been produced.

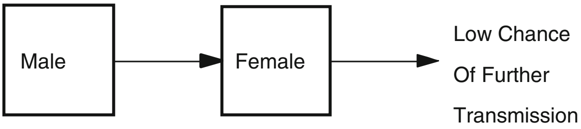
4.1 Why Was the AIDS Epidemic so Much More Prevalent in America Than in Other First World Countries?

Although the popular press in the early 1980s talked of AIDS as being a new disease prudence and experience indicated that it was not. Just as new species of animals have not been noted during human history, the odds for a sudden appearance (absent genetic engineering) of a new virus are not good. My own discussions with pathologists with some years of experience gave anecdotal cases of young Anglo males who had presented with Kaposi's sarcoma at times going back to early days in the pathologists' careers. This pathology, previously seldom seen in persons of Northern European extraction, now widely associated with AIDS, was at the time simply noted as isolated and unexplained. Indeed, a few years after the discovery of the HIV virus, HIV was discovered in decades old refrigerated human blood samples from both Africa and America.

Although it was clear that AIDS was not a new disease, as an epidemic it had never been recorded as such. Because some early cases were from the Congo, there was an assumption by many that the disease might have its origins there. Record keeping in the Congo was not and is not very good. But Belgian colonial troops had been located in that region for many years. Any venereal disease acquired in the Congo should have been vectored into Europe in the 19th century. But no AIDS-like disease had been noted. It would appear, then, that AIDS was not contracted easily as is the case, say, with syphilis. Somehow, the appearance of AIDS as an epidemic in the 1980s, and not previously, might be connected with higher rates of promiscuous sexual activity made possible by the relative affluence of the times.

Then there was the matter of the selective appearance of AIDS in the American homosexual community. If the disease required virus in some quantity for effective transmission (the swift progression of the disease in hemophiliacs plus the lack of notice of AIDS in earlier times gave clues that such might be the case), then the profiles in Figs. 5 and 6 give some idea of why the epidemic seemed to be centered in the American homosexual community. If passive to active transmission is much less likely than active to passive, then clearly the homosexual transmission patterns facilitate the disease more than the heterosexual ones.

Fig. 5 Heterosexual transmission of AIDs



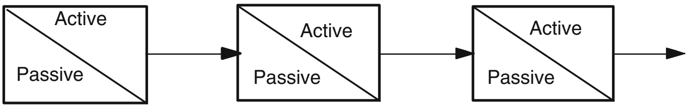


Fig. 6 Homosexual transmission of AIDs

One important consideration that seemed to have escaped attention was the appearance of the epidemic in 1980 instead of 10 years earlier. Gay lifestyles had begun to be tolerated by law enforcement authorities in the major urban centers of America by the late 1960s. If homosexuality was the facilitating behavior of the epidemic, then why no epidemic before 1980? Of course, believers in the "new disease" theory could simply claim that the causative agent was not present until around 1980. In the popular history of the early American AIDS epidemic, *And the Band Played On*, Randy Shilts points at a gay flight attendant from Quebec as a candidate for "patient zero." But this "Lunar Pox" theory was not a position that any responsible epidemiologist could take (and, indeed, as pointed out, later investigations revealed HIV samples in human blood going back into the 1940s).

What accounts for the significant time differential between civil tolerance of homosexual behavior prior to 1970 and the appearance of the AIDS epidemic in the 1980s? Were there some other sociological changes that had taken place in the late 1970s that might have driven the endemic over the epidemiological threshold?

It should be noted that in 1983, data were skimpy and incomplete. As is frequently the case with epidemics, decisions need to be made at the early stages when one needs to work on the basis of skimpy data, analogy with other historical epidemics, and a model constructed on the best information available.

I remember in 1983 thinking back to the earlier American polio epidemic that had produced little in the way of sociological intervention and less in the way of models to explain the progress of the disease. Although polio epidemics had been noted for some years (the first noticed epidemic occurred around the time of World War I in Stockholm), the American public health service had indeed treated it like the "Lunar Pox." That is, they discarded sociological intervention based on past experience of transmission pathways and relied on the appearance of vaccines at any moment. They had been somewhat lucky, since Dr. Jonas Salk started testing his vaccine in 1952 (certainly they were luckier than the thousands who had died and the tens of thousands who had been permanently crippled). But basing policy on hope and virological research was a dangerous policy (how dangerous we are still learning as we face the reality of 650,000 Americans dead by 2011 from AIDS). I am unable to find the official CDC death count in America as of the end of 2014, but a senior statistician colleague from CDC reckons that 700,000 is not unreasonable.

Although some evangelical clergymen inveighed against the epidemic as divine retribution on homosexuals, the function of epidemiologists is to use their God-given wits to stop epidemics. In 1983, virtually nothing was being done except to wait for virological miracles.

One possible candidate was the turning of a blind eye by authorities to the gay bathhouses that started in the late 1970s. These were places where gays could engage in high frequency anonymous sexual contact. By the late 1970s they were allowed to operate without regulation in the major metropolitan centers of America. My initial intuition was that the key was the total average contact rate among the target population. Was the marginal increase in the contact rate facilitated by the bathhouses sufficient to drive the endemic across the epidemiological threshold? It did not seem likely. Reports were that most gays seldom (many, never) frequented the bathhouses.

In the matter of the present AIDS epidemic in the United States, a great deal of money is being spent. However, practically nothing in the way of steps for stopping the transmission of the disease is being done (beyond education in the use of condoms). Indeed, powerful voices in the Congress speak against any sort of government intervention. On April 13, 1982, Congressman Henry Waxman [7] stated in a meeting of his Subcommittee on Health and the Environment, "I intend to fight any effort by anyone at any level to make public health policy regarding Kaposi's sarcoma or any other disease on the basis of his or her personal prejudices regarding other people's sexual preferences or life styles." (It is significant that Representative Waxman has been one of the most strident voices in the fight to stop smoking and global warming, considering rigorous measures acceptable to end these threats to human health.)

In light of Congressman Waxman's warnings, it would have taken brave public health officials to close the gay bathhouses. We recall how Louis Pasteur had been threatened with the guillotine if he insisted on proceeding with his rabies vaccine and people died as a result. He proceeded with the testings, starting on himself. There were no Louis Pasteurs at the CDC. The Centers for Disease Control have broad discretionary powers and its members have military uniforms to indicate their authority. They have no tenure, however. The Director of the CDC could have closed the bathhouses, but that would have been an act of courage which could have ended his career. Of all the players in the United States AIDS epidemic, Congressman Waxman may be more responsible than any other for what has turned out to be a death tally exceeding any of America's wars, including its most lethal, the American War Between the States (aka the Civil War).

5 The Effect of the Gay Bathhouses

But perhaps my intuitions were wrong. Perhaps it was not only the total average contact rate that was important, but a skewing of contact rates, with the presence of a high activity subpopulation (the bathhouse customers) somehow driving the epidemic. It was worth a modeling try.

The model developed in [8] considered the situation in which there are two subpopulations: the majority, less sexually active, and a minority with greater activity than that of the majority. We use the subscript "1" to denote the majority portion of the target (gay) population, and the subscript "2" to denote the minority portion.

The latter subpopulation, constituting fraction p of the target population, will be taken to have a contact rate τ times the rate k of the majority subpopulation. The following differential equations model the growth of the number of susceptibles X_i and infectives Y_i in subpopulation i $(i = 1, 2)$.

$$\begin{aligned}
\frac{dY_1}{dt} &= \frac{k\alpha X_1(Y_1 + \tau Y_2)}{X_1 + Y_1 + \tau(Y_2 + X_2)} - (\gamma + \mu)Y_1,\\
\frac{dY_2}{dt} &= \frac{k\alpha\tau X_2(Y_1 + \tau Y_2)}{X_1 + Y_1 + \tau(Y_2 + X_2)} - (\gamma + \mu)Y_2,\\
\frac{dX_1}{dt} &= -\frac{k\alpha X_1(Y_1 + \tau Y_2)}{X_1 + Y_1 + \tau(Y_2 + X_2)} + (1 - p)\lambda - \mu X_1,\\
\frac{dX_2}{dt} &= -\frac{k\alpha\tau X_2(Y_1 + \tau Y_2)}{X_1 + Y_1 + \tau(Y_2 + X_2)} + p\lambda - \mu X_2.
\end{aligned} \tag{2}$$

where

k = number of contacts per month,
α = probability of contact causing AIDS,
λ = immigration rate into the population,
μ = emigration rate from the population,
γ = marginal emigration rate from the population due
to sickness and death.

In Thompson [8], it was noted that if we started with 1,000 infectives in a target population with $k\alpha = 0.05$, $\tau = 1$, a susceptible population of 3,000,000 and the best guesses then available ($\mu = 1/(15 \times 12) = 0.00556$, $\gamma = 0.1$, $\lambda = 16{,}666$) for the other parameters, the disease advanced as shown in Table 2.

Next, a situation was considered in which the overall contact rate was the same as in Table 2, but it was skewed with the more sexually active subpopulation 2 (of size 10 %) having contact rates 16 times those of the less active population.

Even though the overall average contact rate in Tables 2 and 3 is the same $(k\alpha)_{\text{overall}} = 0.05$, the situation is dramatically different in the two cases. Here, it seemed, was a *prima facie* explanation as to how AIDS was pushed over the

Table 2 Extrapolated AIDS cases: $k\alpha = 0.05$, $\tau = 1$

Year	Cumulative deaths	Fraction infective
1	1751	0.00034
2	2650	0.00018
3	3112	0.00009
4	3349	0.00005
5	3571	0.00002
10	3594	0.000001

Table 3 Extrapolated AIDS cases: $k\alpha = 0.02$, $\tau = 16$, $p = 0.10$

Year	Cumulative deaths	Fraction infective
1	2,184	0.0007
2	6,536	0.0020
3	20,583	0.0067
4	64,157	0.0197
5	170,030	0.0421
10	855,839	0.0229
15	1,056,571	0.0122
20	1,269,362	0.0182

threshold to a full-blown epidemic in the United States: a small but sexually very active subpopulation.

This was the way things stood in 1984 when I presented my AIDS paper at the summer meetings of the Society for Computer Simulation in Vancouver. It hardly created a stir among the mainly pharmacokinetic audience who attended the talk. And, frankly, at the time I did not think too much about it because I supposed that probably even as the paper was being written, the "powers that be" were shutting down the bathhouses. The deaths at the time were numbered in the hundreds, and I did not suppose that things would be allowed to proceed much longer without sociological intervention. Unfortunately, I was mistaken.

In November 1986, the First International Conference on Population Dynamics took place at the University of Mississippi where there were some of the best biomathematical modelers from Europe and the United States. I presented my AIDS results [9], somewhat updated, at a plenary session. By this time, I was already alarmed by the progress of the disease (over 40,000 cases diagnosed and the bathhouses still open). The bottom line of the talk had become more shrill: namely, every month delayed in shutting down the bathhouses in the United States would result in thousands of deaths. The reaction of the audience this time was concern, partly because the prognosis seemed rather chilling, partly because the argument was simple to follow and seemed to lack holes, and partly because it was clear that something was pretty much the matter if things had gone so far off track.

After the talk, the well-known Polish probabilist Robert Bartoszyński, with whom I had carried out a lengthy modeling investigation of breast cancer and melanoma (at the Curie-Sklodowska Institute in Poland and at Rice), took me aside and asked whether I did not feel unsafe making such claims. "Who," I asked, "will these claims make unhappy"? "The homosexuals," said Bartoszyński. "No, Robert," I said, "I am trying to save their lives. It will be the public health establishment who will be offended."

And so it has been in the intervening years. I have given AIDS talks before audiences with significant gay attendance in San Francisco, Houston, Washington, and other locales without any gay person expressing offense. Indeed, in his 1997 book [10], Gabriel Rotello, one of the leaders of the American gay community, not only acknowledges the validity of my model but also constructs a survival plan for gay society in which the bathhouses have no place.

5.1 A More Detailed Look at the Model

A threshold investigation of the two-activity population model (2) is appropriate here. Even today, let alone in the mid-1980s, there was no chance that one would have reliable estimates for all the parameters k, α, γ, μ, λ, p, τ. Happily, one of the techniques sometimes available to the modeler is the opportunity to express the problem in such a form that most of the parameters will cancel out. For the present case, we will attempt to determine the $k\alpha$ value necessary to sustain the epidemic when the number of infectives is very small. For this epidemic in its early stages one can manage to get a picture of the bathhouse effect using only a few parameters: namely, the proportion p of the target population which is sexually very active and the activity multiplier τ.

For $Y_1 = Y_2 = 0$ the equilibrium values for X_1 and X_2 are $(1 - p)(\lambda/\mu)$ and $p(\lambda/\mu)$, respectively. Expanding the right-hand sides of (2) in a Maclaurin series, we have (using lower case symbols for the perturbations from 0)

$$\frac{dy_1}{dt} = \left[\frac{k\alpha(1-p)}{1-p+\tau p} - (\gamma+\mu)\right] y_1 + \frac{k\alpha(1-p)\tau}{1-p+\tau p}\, y_2$$
$$\frac{dy_2}{dt} = \frac{k\alpha\tau p}{1-p+\tau p}\, y_1 + \left[\frac{k\alpha\tau^2 p}{1-p+\tau p} - (\gamma+\mu)\right] y_2.$$

Summing then gives

$$\frac{dy_1}{dt} + \frac{dy_2}{dt} = [k\alpha - (\gamma+\mu)]\, y_1 + [k\alpha\tau - (\gamma+\mu)]\, y_2.$$

In the early stages of the epidemic,

$$\frac{dy_1/dt}{dy_2/dt} = \frac{(1-p)}{p\tau}.$$

That is to say, the new infectives will be generated proportionately to their relative numerosity in the initial susceptible pool times their relative activity levels. So, assuming a negligible number of initial infectives, we have

$$y_1 = \frac{(1-p)}{p\tau} y_2.$$

Substituting in the expression for $dy_1/dt + dy_2/dt$, we see that for the epidemic to be sustained, we must have

$$k\alpha > \frac{(1+\mu)(1-p+\tau p)}{1-p+p\tau^2}(\gamma+\mu). \tag{3}$$

Accordingly we define the *heterogeneous threshold* via

$$k_{\mathrm{het}}\alpha = \frac{(1+\mu)(1-p+\tau p)}{1-p+p\tau^2}(\gamma+\mu).$$

Now, in the homogeneous contact case (i.e., $\tau = 1$), we note that for the epidemic not to be sustained, the condition in Eq. (4) must hold.

$$k\alpha < (\gamma+\mu). \tag{4}$$

Accordingly we define the *homogeneous threshold* by

$$k_{\mathrm{hom}}\alpha = (\gamma+\mu).$$

For the heterogeneous contact case with k_{het}, the average contact rate is given by

$$k_{\mathrm{ave}}\alpha = p\tau(k_{\mathrm{het}}\alpha) + (1-p)(k_{\mathrm{het}}\alpha) = \frac{(1+\mu)(1-p+\tau p)}{1-p+p\tau^2}(\gamma+\mu).$$

Dividing the sustaining value $k_{\mathrm{hom}}\alpha$ by the sustaining value $k_{\mathrm{ave}}\alpha$ for the heterogeneous contact case then produces

$$Q = \frac{1-p+\tau^2 p}{(1-p+\tau p)^2}.$$

Notice that we have been able here to reduce the parameters necessary for consideration from seven to two. This is fairly typical for model-based approaches: the dimensionality of the parameter space may be reducible in answering specific questions. Figure 7 shows a plot of this "enhancement factor" Q as a function of τ. Note that the addition of heterogeneity to the transmission picture has roughly the same effect as if all members of the target population had more than doubled their contact rate. Remember that the picture has been corrected to discount any increase in the overall contact rate which occurred as a result of adding heterogeneity. In other words, the enhancement factor is totally a result of heterogeneity. It is this heterogeneity effect which I have maintained (since 1984) to be the cause of AIDS getting over the threshold of sustainability in the United States. Data from the CDC on AIDS have been other than easy to find. Concerning the first fifteen years of the epidemic, Dr. Rachel MacKenzie of the WHO was kind enough to give me the data. Grateful though I was for that data, I know there was some displeasure from the WHO that she

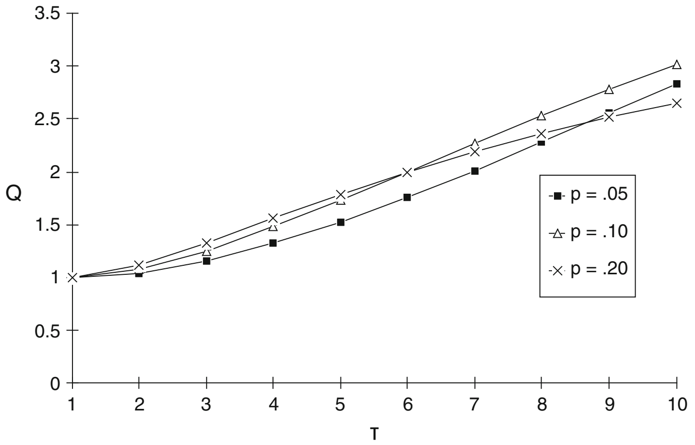


Fig. 7 Effect of a high activity subpopulation

had done so, and after 1995 the data appeared on the internet very irregularly with two and three year gaps between data postings. Since the United States was contributing most of the money for AIDS conferences, grants and other activities, I can understand the reluctance of the WHO to give out information which showed how badly the Americans were doing compared to the rest of the First World. Transparency is generally assumed in scientific research, but that assumption is unfortunately wrong in some of the most important situations. Suffice it to say that during the 15 years of WHO data I was presented, the United States had 10 times the AIDS rate per 100,000 of the UK, 8 times that of Netherlands, 7 times that of Denmark, 4 times that of Canada, and 3.5 times that of France. One can understand the embarrassment of the American CDC. I regret to say that AIDS goes largely unmentioned and unnoticed by the American media and such agencies as the NIH, the PHS, and the NCI. Benjamin Franklin once said: “Experience keeps a hard school and a fool will learn by none other.” What about those who continue failed policies ad infinitum? I believe Albert Einstein called them insane.

Sometimes establishment inertia trumps facts. When I started my crusade against the bathhouses, there were two in Houston. Now, within 5 miles of the Texas Medical Center, there are 17. One of these adjoins the hotel Rice frequently uses to house its visitors. Vancouver, which had no bathhouses when I gave my first AIDS lecture there, now has 3. As some may remember if they attended the recent national meetings of the ASA held in Vancouver, the Gay Pride Parade there has floats from the major Canadian banks and from the University of British Columbia School of Medicine. Gay bathhouses are popping up in several European cities as well. The American AIDS establishment has the pretence of having drugs which can make an AIDS sufferer as treatable as a diabetic. That these drugs are dangerous and over time frequently produce pain so severe tht users eventually opt for cessation of treatment is not much spoken about.

6 Conclusions

Data analysis to a purpose is generally messy. If I think back on the very many consulting jobs I have done over the years, very few were solvavable unless one went outside the box of classical statistical tools into other disciplines and murky waters. Indeed, the honoree of this *Festschrift* Jacek Koronacki is a good example to us all of not taking the easy way out. During martial law, I offered him a tenured post at Rice. I cautioned him that in the unlikely event the Red Army ever left Poland, the next administration would be full of unsavory holdovers from the junior ranks of the Party posing as Jeffersonian reformers. Jacek left Rice, nevertheless, with his wife, daughter and unborn son. He said he could not think of abandoning Poland and his colleagues. It would be ignoble to do so. He would return to Poland with his family and hope God would provide. I have to say that though I was correct in my prophecy, Jacek chose the right path.

References

1. Sharpe, WE (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Finance 19:425–442
2. Sharpe William E (2000) Portfolio theory and capital markets. McGraw Hill, New York
3. Bogle JC (1999) Common sense and mutual funds: new imperatives for the intelligent investor. Wiley, New York
4. Thompson JR, Baggett LS, Wojciechowski WC, Williams EE (2006) Nobels for nonsense. J Post Keynesian Econ Fall 3–18
5. Thompson, JR (2010) Methods and apparatus for determining a return distribution for an investment portfolio. US Patent 7,720,738 B2, 18 May 2010
6. Baggett LS, Thompson JR (2007) Every man's maxmedian rule for portfolio management. In: Proceedings of the 13th army conference on applied statistics
7. Shilts R (1987) And the band played on: politics, people, and the AIDS epidemic. St. Martin's Press, New York, p 144
8. Thompson JR (1984) Deterministic versus stochastic modeling in neoplasia. In: Proceedings of the 1984 computer simulation conference, society for computer simulation, San Diego, 1984, pp 822–825
9. Thompson JR (1998) The united states AIDS epidemic in first world context. In: Arino O, Axelrod D, Kimmel M (eds) Advances in mathematical population dynamics: molecules, cells and man. World Scientific Publishing Company, Singapore, pp 345–354
10. Rotello G (1997) Sexual ecology: AIDS and the destiny of Gay men. Dutton, New York, pp 85−89