Springer Proceedings in Mathematics & Statistics

Jaime A. Londoño
José Garrido
Monique Jeanblanc
Editors

Actuarial Sciences and Quantitative Finance

ICASQF2016, Cartagena, Colombia, June 2016

Springer

Springer Proceedings in Mathematics & Statistics

Volume 214

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Jaime A. Londoño • José Garrido
Monique Jeanblanc
Editors

Actuarial Sciences and Quantitative Finance

ICASQF2016, Cartagena, Colombia, June 2016

Springer

Editors
Jaime A. Londoño
Departamento de Matemáticas y Estadí-stica
Universidad Nacional de Colombia
Manizales, Caldas, Colombia

Department of Mathematics
National University of Colombia
Bogota, Bogota, Colombia

Monique Jeanblanc
LaMME, Batiment IBGBI
Université d'Evry Val D'Essone
Evry Cedex, Essonne, France

José Garrido
Department of Mathematics and Statistics
Concordia University
Montréal, QC, Canada

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-66534-4 ISBN 978-3-319-66536-8 (eBook)
https://doi.org/10.1007/978-3-319-66536-8

Library of Congress Control Number: 2017955004

Mathematics Subject Classification (2010): 62P05, 91B30, 91G20, 91G80

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The chapters in this volume of the Springer Proceedings in Mathematics and Statistics entitled "Actuarial Sciences and Quantitative Finance: ICASQF2016, Cartagena, Colombia, June 2016" are from selected papers presented at the Second International Congress on Actuarial Science and Quantitative Finance, which took place in Cartagena from June 15 to 18, 2016. The conference was organized jointly by the Universidad Nacional de Colombia, Universidad de Cartagena, Universidad del Rosario, Universidad Externado de Colombia, Universidad de los Andes, ENSIIE/Université Evry Val d'Essonne, and ADDACTIS Latina. It also received support from Universidad Industrial de Santander, Ambassade de France en Colombie, and ICETEX. The conference took place in the Claustro de San Agustín and Casa Museo Arte y Cultura la Presentación in the walled city of Cartagena.

This congress was the second edition of a series of events to be organized every other year, with the objective of becoming a reference in actuarial science and quantitative finance in Colombia, the Andean region (Peru, Colombia, Venezuela, Ecuador, and Bolivia), and the Caribbean. The congress had participation from researchers, students, and practitioners from different parts of the world. This second edition helped enhance the relations between the academic and industrial actuarial and financial communities in North America, Europe, and other regions of the world.

The emphasis of the event was equally distributed between actuarial sciences and quantitative finance and covered a variety of topics such as Statistical Techniques in Finance and Actuarial Science, Portfolio Management, Derivative Valuation, Risk Theory, Life and Pension Insurance Mathematics, Non-life Insurance Mathematics, and Economics of Insurance.

The event consisted of plenary sessions with invited speakers in the areas of actuarial science and quantitative finance, oral sessions of contributed talks on these topics, as well as short courses taught by some of the invited speakers and poster sessions. The list of invited speakers reflects the broad variety of topics: Nicole El Karoui (Self-Exciting Process in Finance and Insurance for Credit Risk and Longevity Risk Modeling in Heterogeneous Portfolios), Julien Guyon

(Path-Dependent Volatility), Christian Hipp (Stochastic Control for Insurance: New Problems and Methods), Jean Jacod (Estimation of Volatility in Presence of High Activity Jumps and Noise), Glenn Meyers (Aggressive Backtesting of Stochastic Loss Reserve Models—Where It Leads Us), Michael Sherris (To Borrow or Insure? Long-Term Care Costs and the Impact of Housing), Qihe Tang (Mitigating Extreme Risks Through Securitization), and Fernando Zapatero (Riding the Bubble with Convex Incentives). Topics for short courses included the following: The New Post-crisis Landscape of Derivatives and Fixed Income Activity Under Regulatory Constraints on Credit Risk, Liquidity Risk, and Counterparty Risk (Nicole El Karoui); Stochastic Control for Insurers: What Can We Learn from Finance, and What Are the Differences? (Christian Hipp); High-Frequency Statistics in Finance (Jean Jacod); and Using Bayesian MCMC Models for Stochastic Loss Reserving (Glenn Meyers).

Additionally, researchers and students presented oral contributions and posters. There were 30 contributed oral presentations, 26 invited oral contributions, and ten poster presentations. We received 85 contributions and 34 invited contributions. The selection process was the result of careful deliberations, and 54 oral contributed presentations of the 85 submissions and 20 posters were accepted. Authors came from different corners of the world and countries of origin including Australia, Brazil, Canada, Chile, Colombia, Egypt, France, Germany, Italy, Jamaica, Mexico, Spain, Switzerland, the United Kindom, Uruguay, and the United States. The number of contributions along with the total number of 279 registered participants shows the steady growth of the congress and its consolidation as the main event of the area in the Andean region and the Caribbean.

The congress put the emphasis on enhancing relations between industry and academia providing a day to address problems arising from the financial and insurance industries. As a matter of fact, topics and speakers themselves came from these sectors. The congress provided practitioners a platform to present and discuss with academics and students different approaches in addressing problems arising from the industries in the region.

The current proceedings are based on invitations to selected oral contributions and selected contributions presented by the invited speakers. All contributions were subject to an additional review process. The spectrum of the eight papers published here reflects the diverse nature of the presentations: there are five papers on actuarial sciences and three papers on quantitative finance.

Special thanks go to the members of the organizing committee, which included Javier Aparicio (Colombia, ADDACTIS Latina), Prof. Sergio Andrés Cabrales (Colombia, Universidad de los Andes), Prof. Carlos Alberto Castro (Colombia, Universidad del Rosario), Prof. Margaret Johanna Garzón (Colombia, Universidad Nacional de Colombia, Bogotá), Prof. Sandra Gutiérrez (Colombia, Universidad de Cartagena), Prof. Jaime A. Londoño (Colombia, Universidad Nacional de Colombia, Bogotá), Prof. Sergio Pulido (France, ENSIIE/Université Evry Val d'Essonne), Prof. Javier Sandoval (Colombia, Universidad Externado de Colombia),

and Prof. Arunachalam Viswanathan (Colombia, Universidad Nacional de Colombia, Bogotá). Finally, we would like to thank all the conference participants who made this event a great success.

Manizales, Colombia — Jaime A. Londoño
Montréal, QC, Canada — José Garrido
Evry Cedex, France — Monique Jeanblanc
May 2017

Contents

Part I
Actuarial Sciences

Robust Paradigm Applied to Parameter Reduction in Actuarial Triangle Models

Gary Venter

Abstract The recognition that models are approximations used to illuminate features of more complex processes brings a challenge to standard statistical testing, which assumes the data is generated from the model. Out-of-sample tests are a response. In my view this is a fundamental change in statistics that renders both classical and Bayesian approaches outmoded, and I am calling it the "robust paradigm" to signify this change. In this context, models need to be robust to samples that are never fully representative of the process. Actuarial models of loss development and mortality triangles are often over-parameterized, and formal parameter-reduction methods are applied to them here within the context of the robust paradigm.

Keywords Loss reserving • Mortality • Bayesian shrinkage • MCMC

1 Introduction

Section 2 discusses model testing under the robust paradigm, including out-of-sample tests and counting the effective number of parameters. Section 3 introduces parameter-reduction methods including Bayesian versions. Section 4 reviews actuarial triangle modeling based on discrete parameters by row, column, etc., and how parameter-reduction can be used for them. Section 5 gives a mortality model example, while Sect. 6 illustrates examples in loss reserving. Section 7 concludes.

2 Model Testing Within the Robust Paradigm

Both Bayesian and classical statistics typically assume that the data being used to estimate a model has been generated by the process that the model specifies. In many, perhaps most, financial models this is not the case. The data is known to come

G. Venter (✉)
University of New South Wales, Sydney, NSW, Australia
e-mail: gary.venter@gmail.com

J.A. Londoño et al. (eds.), *Actuarial Sciences and Quantitative Finance*, Springer
Proceedings in Mathematics & Statistics 214, https://doi.org/10.1007/978-3-319-66536-8_1

from a more complex process and the model is a hopefully useful but simplified representation of that process. Goodness-of-fit measures that assume that the data has been generated from the sample are often not so reliable in this situation, and out-of-sample tests of some sort or another are preferred. These can help address how well the model might work on data that was generated from a different aspect of the process. I have coined the term "robust paradigm" to refer to statistical methods useful when the data does not come from the model.

Much statistics today is based on pragmatic approaches that keep the utility of the model for its intended application in mind, and regularly deviate from both pure Bayesian and pure classical paradigms. That in itself does not mean that they are dealing with data that does not come from the models. In fact, even out-of-sample testing may be done purely to address issues of sample bias in the parameters, even assuming that the data did come from the model. But simplified models for complex processes are common and pragmatic approaches are used to test them. This is what is included in the robust paradigm.

When models are simplified descriptions of more complex processes, you can never be confident that new data from the same process will be consistent with the model. In fact with financial data, it is not unusual for new data to show somewhat different patterns from those seen previously. However, if the model is robust to a degree of data change, it may still work fairly well in this context. More parsimonious models often hold up better when data is changing like that. Out-of-sample testing methods are used to test for such robustness.

A typical ad hoc approach is the rotating $\frac{4}{5}$ths method: the data is divided, perhaps randomly, into five subsets, and the model is fit to every group of four of these five. Then the fits are tested on the omitted populations, for example by computing the negative loglikelihood (NLL). Competing models can be compared on how well they do on the omitted values.

A well-regarded out-of-sample test is leave one out, or "loo." This fits the model many times, leaving out one data point at a time. Then the fit is tested at each omitted point to compare alternative models. The drawback is in doing so many fits for each model.

In Bayesian estimation, particularly in Markov Chain Monte Carlo (MCMC), there is a shortcut to loo. The estimation produces many sample parameter sets from the posterior distribution of the parameters. By giving more weight to the parameter sets that fit poorly at a given data observation, an approximation to the parameters that would be obtained without that observation can be made. This idea is not new, but such approximations have been unstable.

A recent advance, called Pareto smoothed importance sampling, appears to have largely solved the instability problem. A package to do this, called `loo`, is available with the `Stan` package for MCMC. It can be used with MCMC estimation not done in `Stan` as well. It allows comparison of the NLL of the omitted points across models. This modestly increases the estimation time, but is a substantial improvement over multiple re-estimation. Having such a tool available makes loo likely to become a standard out-of-sample fitting test.

This is a direct method to test models for too many parameters. Over-fitted models will not perform well out of sample. If the parameters do better out of sample, they are worth it. Classical methods for adjusting for over-parameterization, like penalized likelihood, are more artificial by comparison, and never have become completely standardized. In classical nonlinear models, counting the effective number of parameters is also a bit complex.

2.1 Counting Parameters

In nonlinear models it is not always apparent how many degrees of freedom are being used up by the parameter estimation. One degree of freedom per parameter is not always realistic, as the form of the model may constrict the ability of parameters to pull the fitted values towards the actual values.

A method that seems to work well within this context is the generalized degrees of freedom method of Ye (1998). Key to this is the derivative of a fitted point from a model with respect to the actual point. That is the degree to which the fitted point will change in response to a change in the actual point. Unfortunately this usually has to be estimated numerically for each data point.

The generalized degrees of freedom of a model fit to a data set is then the sum across all the data points of the derivatives of the fitted points with respect to the actual points, done one at a time. In a linear model this is just the number of parameters. It seems to be a reasonable representation of the degrees of freedom used up by a model fit, and so can be used like the number of parameters is used in linear models to adjust goodness-of-fit measures, like NLL. A method of counting the effective number of parameters is also built into the `loo` package.

3 Introduction to Parameter Reduction Methodology

Two currently popular parameter reduction methodologies are:

- Linear mixed models (LMM), or in the GLM environment GLMM
- Lasso—Least Absolute Shrinkage and Selection Operator

3.1 Linear Mixed Models

LMM starts by dividing the explanatory variables from a regression model into two types: fixed effects and random effects. The parameters of the random effects are to be shrunk towards zero, based perhaps on there being some question about whether or not these parameters should be taken at face value. See for example Lindstrom and Bates (1990) for a discussion in a more typical statistical context.

Suppose you are doing a regression to estimate the contribution of various factors to accident frequency of driver/vehicle combinations. You might make color of car a random effect, thinking that probably most colors would not affect accident frequency, but a few might, and even for those you would want the evidence to be fairly strong. Then all the parameters for the car color random effects would be shrunk towards or to zero, in line with this skepticism but with an openness to being convinced.

This could be looked at as an analysis of the residuals. Suppose you have done the regression without car color but suspect some colors might be important. You could divide the residuals into groups by car color. Many of these groups of residuals might average to zero, but a few could have positive or negative mean—some of those by chance, however. In LMM you give color i parameter b_i and specify that b_i is normal with mean zero and variance $d_i\sigma^2$, where σ^2 is the regression variance and d_i is a variance parameter for color i. LMM packages like in SAS, Matlab, R, etc. generally allow a wide choice of covariance matrices for these variances, but we will mainly describe the base case, where all of them are independent.

The d_i's are also parameters to be estimated. A color with consistently high residuals is believably a real effect, and it would be estimated with a fairly high d_i to allow b_i to be away from zero. The b_i's are usually assumed to be independent of the residuals. LMM simultaneously maximizes the probability of the b_i's, $P(b)$, and the conditional probability of the observations given b, $P(y|b)$, by maximizing the joint likelihood $P(y, b) = P(y|b)P(b)$.

For a b_i parameter to get further from zero, it has to improve the likelihood of the data by more than it hurts the density of the b's. This is more likely if the previous residuals for that color are grouped more tightly around their mean, in the color example. Then the parameter would help the fit for all those observations. That clustering is not what is measured by d_i, however. It instead determines how much b_i could differ from zero, and its estimate increases to accommodate a useful b_i.

3.2 Lasso

Lasso is a regression approach that constrains the sum of the absolute values of the parameters. It is related to ridge regression, which limits the sum of squares of the parameters. In practice with a lot of variables included, Lasso actually shrinks a fair number of the parameters to zero, eliminating them from the model, whereas ridge regression ends up with many small parameters near zero. Lasso is preferred by most modelers for this reason, and is also preferable to stepwise regression.

In its standard application, all the parameters except the constant term are shrunk, although there is no reason some parameters could not be treated like fixed effects and not shrunk. See Osbourne et al. (2000) for an introduction. Also Pereira et al. (2016) gives examples more general than standard regression.

To make the competition among the independent variables fair, all of them are standardized to have mean zero and variance one by subtracting a constant and

dividing by a constant. The additive transform gets built into the constant term of the regression, and the multiplicative one scales the parameter of that variable.

Then what is minimized is the NLL plus a selected factor times the sum of the absolute values of the parameters. The selection of the factor can be subjective—several are tried with the resulting models evaluated by expert judgment. Using loo to see how well the models with different factors do on the omitted points is more highly regarded, but in a classical setting requires a lot of re-estimation, depending on the sample size.

3.3 Problem with LMM: All Those Variances

Counting parameters is an issue with classical Lasso and LMM. For both, fewer degrees of freedom are used than the apparent number of parameters, due to the constraints. For LMM there is a partial shortcut to counting parameters.

In a regression, the so-called hat matrix is an $N \times N$ matrix, where N is the sample size, which can be calculated from the matrix of independent variables—the design matrix. Multiplying the hat matrix on the right by the vector of observations gives the vector of fitted values. The diagonal of the hat matrix thus gives the response of a fitted value to its observation, and in fact is the derivative of the fitted value with respect to the actual value.

The sum of the diagonal of the hat matrix is thus the generalized degrees of freedom. This holds in LMM as well, but only conditional on the estimated variances. Thus the degrees of freedom used up in estimating the variances do not show up in the hat matrix.

Different LMM estimation platforms can give slightly different parameters—but usually with fits of comparable quality. One triangle model we fit, similar to those discussed below, nominally had 70 parameters, not including the variances. We fit it with two methods. Using the diagonal of the hat matrix indicated that 17.3 degrees of freedom were used by one fitting method, and 19.9 by the other. The second one had a slightly lower NLL, and the penalized likelihoods, by any methods, were comparable.

Since these parameter counts are conditional on the estimated variances d_i, we then did a grind-out generalized-degrees-of-freedom calculation by re-estimating the model changing each observation slightly, one at a time. That got the variances into the parameter counts. The same two methods as before yielded 45.1 and 50.7 degrees of freedom used, respectively. That means that 27.8 and 30.8 degrees of freedom, respectively, were used up in estimating the variances.

In essence, the fitted values responded much more to changes in the actual values than you would have thought from the hat matrix. The parameter reduction from the apparent 70 original parameters was much less than it at-first appeared to be. For the models we were fitting we concluded that base LMM with variances estimated for each parameter was not as effective at parameter reduction as we had thought. This lends more support to using Lasso, or perhaps LMM with fewer, perhaps just a single, variance to estimate. That is, you could assume the d_i are all the same.

3.4 Bayesian Parameter Reduction

A way to shrink parameters towards zero in Bayesian models is to use shrinkage priors. These are priors with mean zero and fairly low variances, so tend to prioritize smaller values of the parameters. An example is the Laplace, or double exponential, distribution, which is exponential in x for $x > 0$ and in $-x$ for $x < 0$:

$$x > 0 : f(x) = e^{-x/b}/2b \tag{1}$$

$$x < 0 : f(x) = e^{x/b}/2b \tag{2}$$

This has heavier tails and more weight near zero than the normal has. Even more so is the horseshoe distribution, which is a normal with σ^2 mixed by a Cauchy.

Typically shrinkage priors are used in MCMC estimation (Fig. 1). There is a lot of flexibility available in the choice of the variances. They can all be the same,

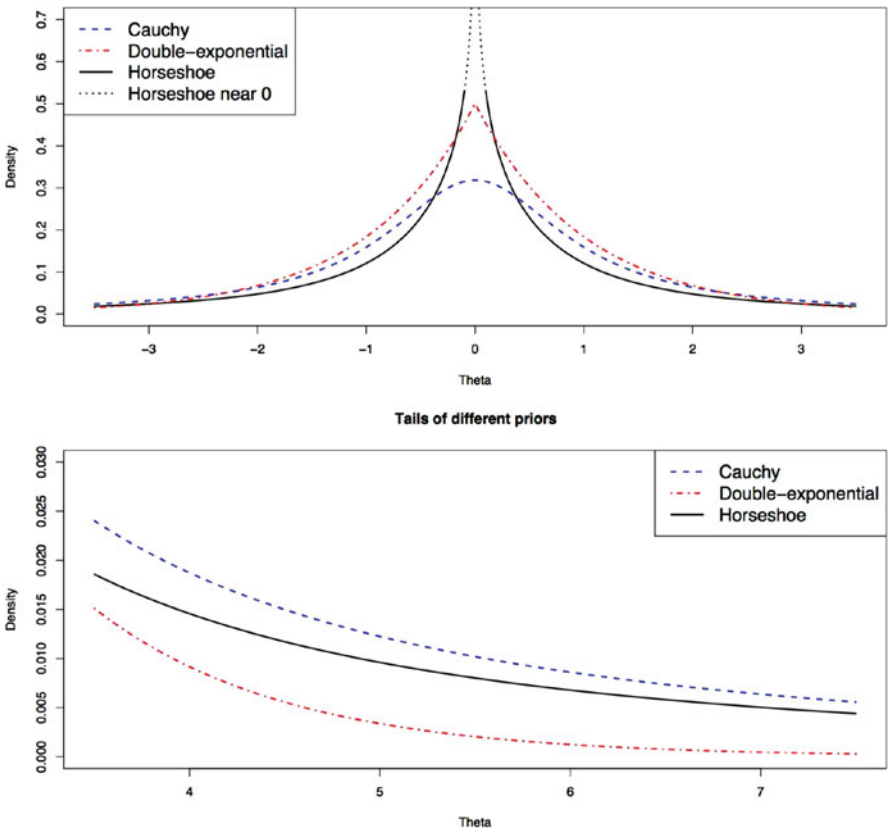


Fig. 1 Shrinkage priors

which is Lasso-like, or vary for different parameters. Some or all of the parameters can have shrinkage priors. Thus the distinctions between LMM and Lasso are not so meaningful in MCMC. There is a wide variety of approaches that can be used.

One fairly viable approach is to use the same variance in the shrinkage priors for all the parameters, and then use loo to see approximately what this variance should be to get the best out-of-sample performance.

3.5 Non-informative Priors

For parameters you do not want to shrink, if you have information or beliefs about a reasonable range for the parameter, that can be coded into the prior distribution. A convenient alternative is non-informative priors. For instance in `Stan`, for a parameter that could be positive or negative, if a prior is not specified the prior is assumed to be uniform on the real line.

This prior density is infinitesimal everywhere and in fact is just specified as being proportional to 1. In `Stan` it is typical to omit constants of proportionality, even if they are not real numbers. This prior, however, viewed as a prior belief, is patently absurd. Most of the probability would lay outside of any finite interval, so it is like saying the parameter probably has a very high absolute value, but we don't know if it is positive or negative.

Nonetheless using it as a prior tends to work out well. Posterior variances from it are often quite similar to what classical statistics would give for estimation variances. Thus the results seem familiar and reasonable. In essence, the prior ceases to be an opinion about the parameter, and instead is chosen because it tends to work out well. This is further evidence that we are no longer in the realm of either classical or Bayesian statistics—it is a pragmatic focus more than a theoretical one.

Things get more awkward when a parameter has to be positive. Assuming uniformity on the positive reals is problematic. While the uniform on the real line has infinite pulls both up and down, on the positive reals the infinite side is only an upward pull. There is thus a tendency for this prior to give a higher estimate than classical statistics would give.

An alternative is to use a prior proportional to $1/x$. This diverges at zero and infinity, so pulls infinitely in both directions. It tends to produce estimates similar to classical unbiased estimates. It is equivalent to giving the log of the parameter a uniform distribution on the reals, which is the easiest way to set it up in `Stan`.

People who do not like non-informative priors sometimes use very diffuse proper priors. One example can be written Gamma(0.001, 0.001). It has mean one and standard deviation about $31\frac{5}{8}$. It is, however, a quite strange prior. Even though the mean is one, the median is in the vicinity of 10^{-300}. The 99th percentile is about 0.025, while the 99.9th is 264 and the 99.99th is 1502. Thus it strongly favors very low values, with occasional very high values showing up. It usually works out alright in the end but can cause difficulty in the estimation along the way.

4 Actuarial Triangle Models with Time Variables

Data for the evolution of insurance liabilities and for mortality can be arranged in two-dimensional arrays, for example with rows for year of origin, and columns for lag to extinction. Actually the time periods are not always years—they could be quarters, months, or even days—but here we will call them years for simplicity. For liabilities, year of origin is often the year the event happened, and lag is the time it takes to close the case and make final payments. For mortality, year of origin is year of birth and lag is the number of years lived. For mortality, the rows are sometimes taken as the calendar years that the extinctions occur in, which is just a different arrangement of the same data—the diagonals are rotated to become the rows, and vice versa.

A common arrangement within the array has the data all above the SW—NE diagonal, giving the term triangle, but various shapes are possible. Mortality triangles for a population usually contain the ratio of deaths in the year to the number alive at the start of the year. Liability triangle cells could contain incremental or cumulative claims payments or claims-department estimates of eventual payments. Here we will assume they are incremental paid losses and are positive or blank.

A popular class of models the log of each entry as the sum of a row effect and a column effect—so there is a dummy variable and a parameter for each row and each column. It is also not unusual to have a parameter for each calendar year, which is the year of origin plus the lag (assuming beginning at lag zero). The calendar-year effects are trends—perhaps inflation for liabilities and increased longevity over time for mortality. It is fairly common in mortality modeling to allow for different ages to feel the longevity improvement more strongly, so an additional parameter might be added for each age as a multiplier to the trend to reflect how strongly that age benefits from the trend.

In doing this modeling actuaries have found that trends in longevity sometimes affect different ages differently, so a single pattern of age-responsiveness does not always hold. To account for this, models now allow a few calendar-year trends, each with its own impact by age. Some models also allow for interaction of age with the year-of-birth cohort parameters, but this effect does not seem to be consistent across cohorts and is less common in recent models. Even in the liability models there could be changes in the lag effects over time, which could be modeled by interactions of lag with year of origin or calendar year.

Letting $p[n]$ be the year-of-origin parameter for year n, $q[u]$ be the age parameter for age u, r refer to a calendar year trend, and s be a set of age weights, the model for the logged value in the n, u cell can be expressed as:

$$y[n,u] = p[n] + q[u] + \sum_i r_i[n+u]s_i[u] + \varepsilon_{n,u} \tag{3}$$

The sum is over the various trends. With a single trend and no age interaction with trend, this would be a typical liability emergence model. There it is not unusual to even leave off the trend entirely—for instance if the trend is constant it will project onto the other two directions.

4.1 Parameter Reduction

The model as stated so far is over-parameterized. One approach to parameter reduction is to require that nearby years or lags have similar parameters. Life insurance actuaries have tried using cubic splines for this. General insurance actuaries have independently been using linear splines. That is, differences between adjacent parameters (i.e., slopes) are constant for a while, with occasional changes in slope. The slope changes are thus the second differences across the parameters.

As the second differences change only occasionally, they are good candidates for parameter-reduction methods. That is the approach explored here. The slope changes are the parameters modeled with specified priors, and these accumulate to the slopes and those to levels, which are the p, q, r, s in the model equation. This can apply to long or short trend periods so can be used for both the life and the general insurance models.

The fitting was done with the `Stan` package, taking double exponential priors for the slope changes. A single variance was specified for all these priors, which in the end was determined by loo in the mortality example. Judgment was used for this in the liability example, but that is not a finished model.

5 Mortality Model Example

US mortality data before 1970 is considered of poor quality, so we use mortality rates in years 1970–2013. Cohorts 1890–1989 were modeled for ages 15–89. A model using Eq. (3) with two trends r_1 and r_2 was selected (i takes on two values: 1 and 2). The first trend is for all the years and the second is zero except for the years 1985–1996, which had increased mortality at younger ages, primarily associated with HIV, but also drug wars. The latter trend was strongest for ages 27–48, so weights were estimated for those years. Here n is the year of birth and u is the age at death, so $n + u$ is the year of death.

The model was calibrated using the MCMC package `Stan` with the second differences of the p, q, r, and s parameters given double exponential priors. Then the parameters in (3) are cumulative sums of the second differences. A lot of the second differences shrink towards zero due to this prior, so the parameters come out looking like they fall on fairly smooth curves—which are actually linear splines.

It is possible to get fairly different parameter sets with quite similar fits, so a fair number of constraints are needed for the sake of specificity. For symmetry, a constant term was added to the model, and then a base mortality parameter q, a trend parameter r, and a cohort parameter p were set to zero. The HIV trend was forced to be upward (positive), and all the trend weights were forced to be in [0,1].

It is a bit awkward in `Stan` to force these parameters to be positive. They are sums of the underlying slope parameters, which in turn are sums of slope changes. Any of those could be negative, as slopes could go up or down. Simple constraints,

like using the minimum of the parameter and zero, are problematic in `Stan` because you then lose derivatives needed for the internal calculations. Squaring the value is awkward as well, as then different paths for the slope changes can get to the same level parameter, which makes it look like the slope changes did not converge. In the end, however, this choice is easier to deal with and was taken. Modeling the logs of the levels as piecewise linear is an alternative worth exploring.

The weights were made to stay in [0,1] by dividing them all by the highest of them after squaring. This may make finding parameters more difficult as well, and it seems to slow down the estimation considerably, but it looks like the best way to get specificity.

Cohort levels are regarded as the year-of-birth effects left after everything else is modeled, so were forced to have zero trend—just by making them the residuals around any trend that did appear.

Another problem with cohorts is that the most recent ones are seen only at young ages, which creates a possible offset with the trend change weights. In fact, giving the most recent cohorts high mortality and simultaneously giving the youngest ages high trend weights gave fairly good fits, but does not seem to be actually occurring.

In the end we forced all cohorts from 1973 to 1989 to have the same parameter—which in fact was made zero to avoid overlap with the constant term. For similar reasons, cohorts 1890–1894 all got the same parameter.

`Stan` simulates parameter sets from the posterior distribution of the parameters in several parallel chains—typically four of them. One check of convergence is to compare the means of each parameter across the chains, and the within and between variances. With these constraints, even though estimation was slow, all the chains had very comparable mean values for every level parameter. The slopes and slope changes from different chains sometimes look like mirror images, however, even though they have the same squares.

The parameters graphed here include all four chains as separate lines mainly to show how well they have converged, as the four lines are all very close.

The main trend is fairly steady improvement, but with a slowdown in the 1990s that is not fully accounted for by the HIV trend, and another slowdown in the last 3–4 years. The trend take-up factors by age range from 65% to 100%, and are lowest in the early 30s and the late 80s (Fig. 2).

The HIV trend is highest in the mid-1990s just before treatments became available (Fig. 3). The ages most affected are the 30s (Fig. 4).

The cohort parameters show a fair degree of variation over time (Fig. 5). Relative to trend, etc. the most longevity is seen in those born in the 1940s and before 1910, with a dip around 1970 as well. While thorough modeling of these patterns is a future project, some clues are available in demographic, macroeconomic, and public health events (Fig. 6).

Those born in 1900 were 70 by the start of this data. The portion of this group that got that old seems to have been particularly hardy. In fact they displayed as much chance to get from age 70 to 90 as those born decades later. The cohort parameters would reflect only the ages in the dataset, so are not necessarily indicative of the cohort mortality for earlier ages.

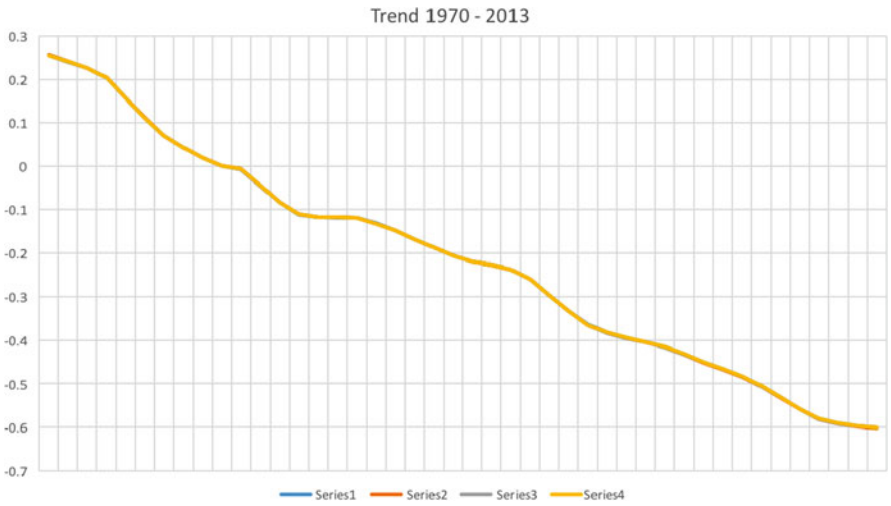


Fig. 2 Time trend 1970–2013

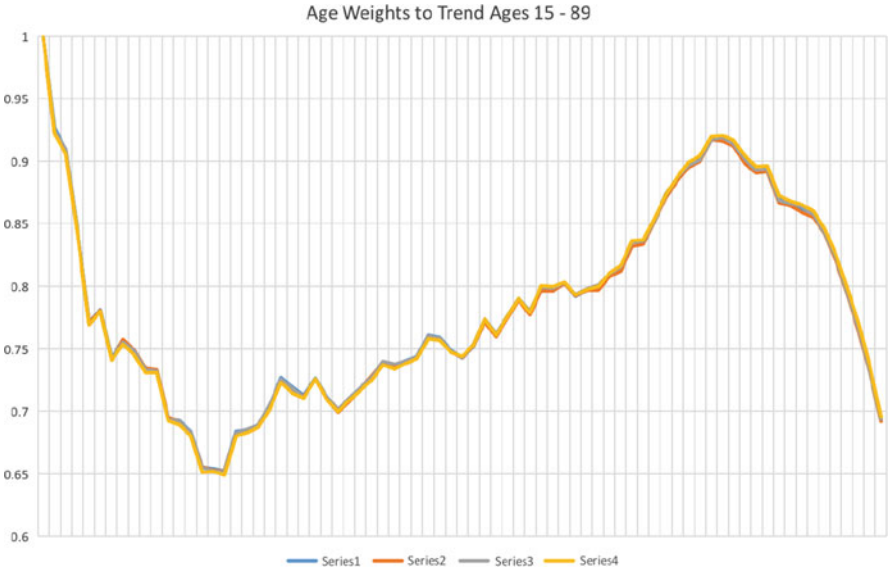


Fig. 3 Age weights to trend ages 15–89

The group born in the 1930s and early 1940s is called the silent generation, or sometimes the fortunate few, and is a unique population. They have had by far the highest real income and net worth of any American generation. This is often attributed to demographics—it was a relatively small population and had little

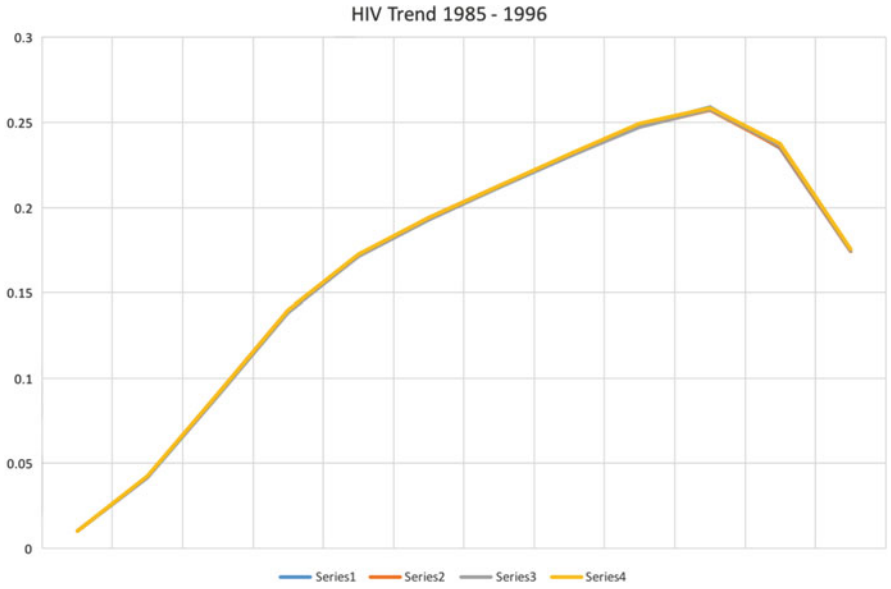


Fig. 4 HIV trend 1985–1996

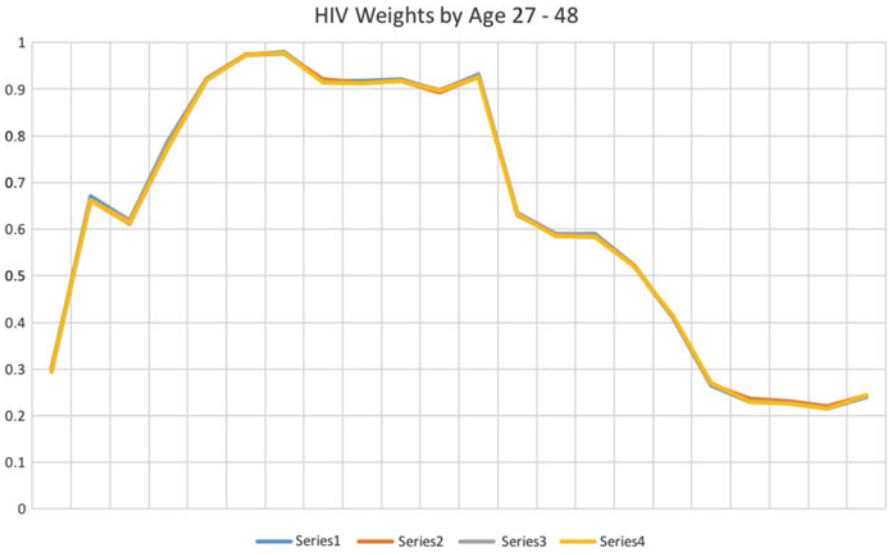


Fig. 5 Age weights to HIV trend ages 27–48

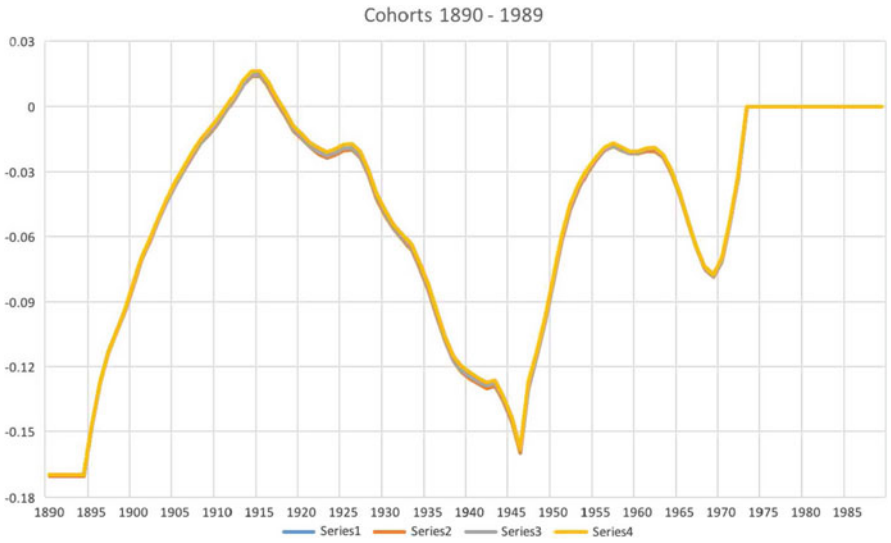


Fig. 6 Cohort level parameters for years of birth 1890–1989

workplace competition from earlier generations. Wealth is linked to longevity and if that were the entire story, this set of cohorts would have had the lowest mortality rates.

However, this was also a generation of heavy smokers. The early boomers, born in the late 1940s, probably smoked less, and had some of the demographic advantages of the fortunate few. The early-boomer cohort may also have been a bit less exposed to obesity than the next group.

Having a small or shrinking population five or so years older seems to be good for career opportunities. Being from the mid-1940s group, I can say that many in my cohort stepped into easily available leadership roles, and hung in there for 30–40 years. The mid-50s cohorts were always back there one level lower—although individual exceptions abound.

The cohorts around 1970 were part of a slowing of population growth that probably also lead to ample career opportunities. Another determinant of career wealth accumulation and so average mortality is the state of the economy upon entering the workforce. That would be another factor to include in this study (Fig. 7).

Looking at the raw mortality rates by age (across) and cohort (down) shows how the age pattern of mortality has been evolving. The width of that graph at an age shows how much mortality improvement that age has experienced from 1970 to 2013.

One thing that stands out is the clumping of lines at the upper right. For most of this period there was little change in the mortality rates at older ages. Then in the last 10 or 11 years, mortality in this group started reducing considerably. This looks

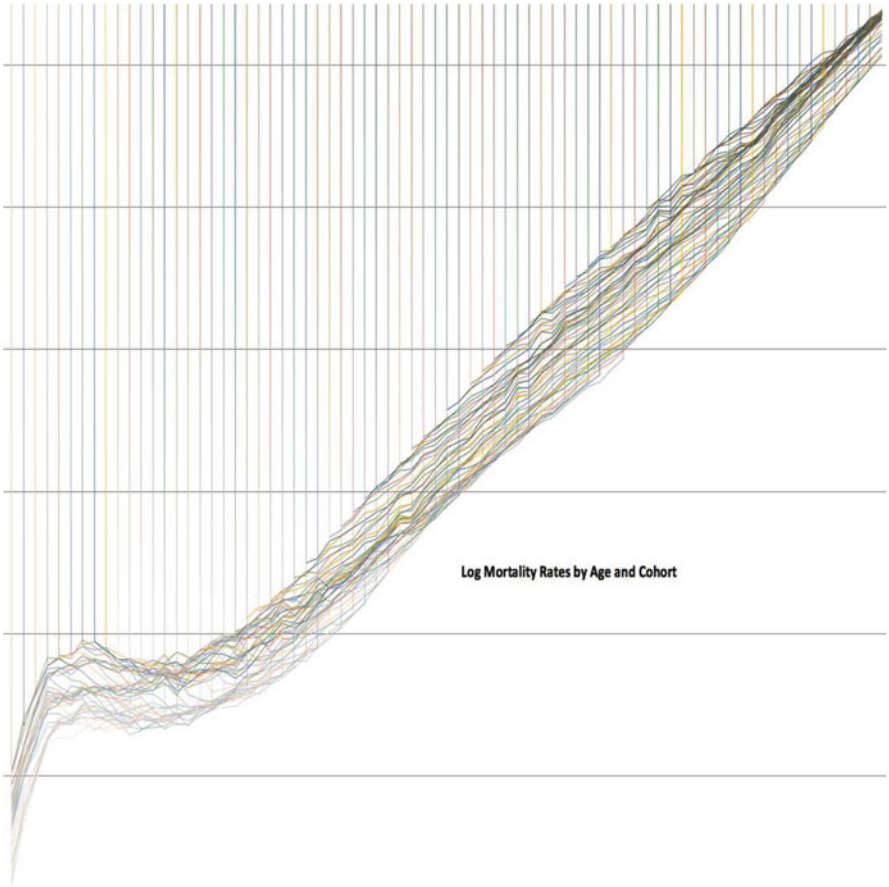


Fig. 7 Log mortality rates by age (increasing from left to right) and cohort (individual lines, most recent generally lower)

like another candidate for a separate trend. Probably the way to do this is to have a separate upward trend in mortality for ages 75+ before 2002, and then give this group the overall trend after that.

Another new trend since 2000 or so is to find little or no improvement in mortality rates for ages in the late 40s through early 60s. This shows up as a clumping of lines at the bottom of the graph above the word "Log." This actually is producing higher mortality for some parts of the population, as has been reported widely in the press. (Our data does not have subpopulation breakouts.) It is again a candidate for its own trend. However, this is also the mid-to-late boomer cohort, which shows up having higher mortality rates anyway, and was also impacted by HIV, so there could be a combination of effects here. Nonetheless, the cohort effect is supposed to be after all other trends have been accounted for, so it seems appropriate to put in a trend here and see what it does to the cohorts.

6 Reserve Modeling Example

Loss reserving has much smaller triangles than mortality does—usually—and simpler models—only one trend and no trend weights by lag typically.

$$y[n,u] = p[n] + q[u] + r[n+u] + \varepsilon_{n,u} \tag{4}$$

We explore here a bit broader model, but will start off with the above. Below is a worker's compensation loss paid triangle for a New Jersey insurer from Taylor and McGuire (2016). The cells are incremental payments.

		0	1	2	3	4	5	6	7	8	9
1988	1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
1989	2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
1990	3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		
1991	4	57,251	49,510	27,036	20,871	14,304	10,552	7,742			
1992	5	59,213	54,129	29,566	22,484	14,114	10,000				
1993	6	59,475	52,076	26,836	22,332	14,756					
1994	7	65,607	44,648	27,062	22,655						
1995	8	56,748	39,315	26,748							
1996	9	52,212	40,030								
1997	10	43,962									

6.1 *Exploratory Analysis*

Looking at residuals from standard development factor analysis can provide information about possible changes in trend and payout patterns. The first test is to calculate the incremental/previous cumulative development factors for each cell, then subtract the column averages from the cell values.

Looking at the results by diagonal can show calendar-year differences. Consistently high or low differences of individual trend factors from column averages along a given diagonal would suggest a possible cost difference for that diagonal compared to the triangle as a whole. It is easier to see such patterns by rotating the triangle so that the diagonals become rows. That was done below with some color coding, and decimals expressed as percents.

	1	2	3	4	5	6	7	8
1	2%							
2	1%	1%						
3	12%	7%	4%					
4	6%	5%	3%	13%				
5	12%	-3%	-11%	-6%	-14%			
6	7%	-0.03%	-1%	-11%	-6%	-5%		
7	-16%	-8%	-0.48%	5%	2%	-1%	-8%	
8	-15%	-6%	2%	-3%	13%	-8%	2%	-0.2%
9	-6%	7%	4%	4%	0.4%	12%	4%	0.2%

It is apparent that the first four diagonals are all positive and the next four mostly negative, with the last again positive. This is suggestive of a calendar-year trend change. The first column seems to be on its own path, however, and may be a payout-change indicator.

A look at payout patterns can be taken by developing each row to ultimate by development factors, then taking the ratio of paid in column to ultimate paid in row for each cell. This can be done for lag zero as well. This test can show changes in payout pattern, but changes in the later columns would be included in averages below that, obfuscating some of the impact.

	0	1	2	3	4	5	6	7	8
1	29%	24%	14%	11%	8%	4%	3%	3%	2%
2	29%	24%	15%	11%	7%	4%	3%	3%	2%
3	28%	26%	15%	10%	6%	5%	3%	3%	
4	28%	25%	13%	10%	7%	5%	4%		
5	28%	26%	14%	11%	7%	5%			
6	29%	25%	13%	11%	7%				
7	32%	22%	13%	11%					
8	31%	21%	15%						
9	30%	23%							

Starting with row 5, there is an increasing trend in payouts at lag 0, offset by a decreasing trend at lag 1. These might reverse slightly in row 9, but that could be due to calendar-year trend.

6.2 Modeling

The model without interaction terms does not include any provision for payout pattern changes. We start with that, however, to see what it says about calendar-year trends, and to see if those could account for the apparent payout shift. Again the double exponential distribution was used for the changes in slope, here with a fairly high variance to make sure that shrinkage was not obscuring any real effects. The development year and accident year parameters came out fairly smooth anyway (Fig. 8).

The main effect seen in the calendar-year trend is a substantial downward jump in 1993. There are two inflation drivers in workers comp. Wage replacement is driven by wage inflation, but is mostly fixed at the wages at time of injury, so shows up in the accident-year, i.e., row, parameters. Medical payments are made at the cost levels at time of payment, on the other hand, so are calendar-year effects (Fig. 9).

Many state laws specify that payments are to be made at the medical providers' standard rates. At some point providers and medical insurers agreed that the providers would increase their rates substantially but those insurers would get a

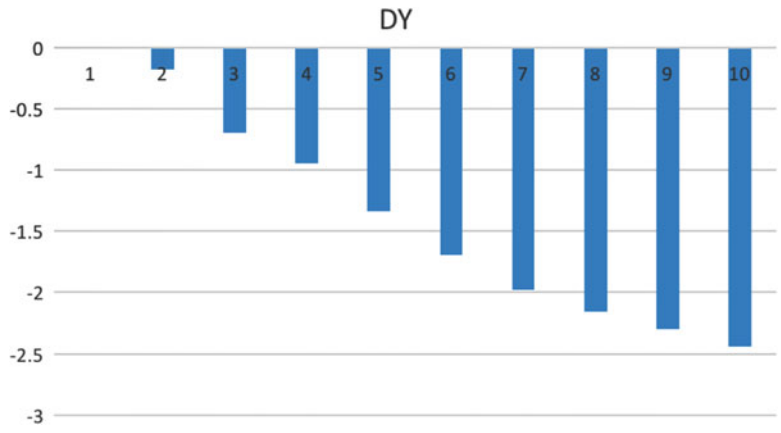


Fig. 8 Log payout level by lag

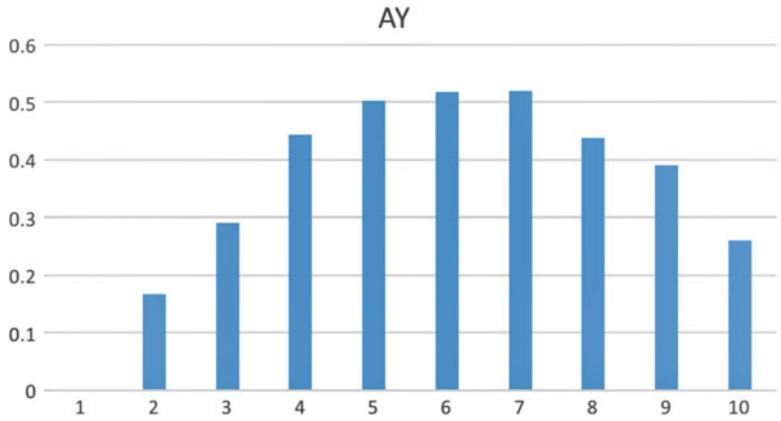


Fig. 9 Accident year levels

discounted rate. That left comp insurers as the only ones paying those artificial standard rates. At some point states started to realize this and basically get the comp insurers inside the game—perhaps through medical fee schedules for comp or other approaches. The comp insurers did not have the political clout to accomplish this, but they pass costs on to employers, who often do. Still, however, some states have higher medical payments for workers comp compared to other insurers (Fig. 10).

The downward jump in costs on the 1993 diagonal could well have come from this kind of reform. By 1997 it appears to be eroding a bit, however.

In any case, this model does not resolve the payout pattern issue. Lag 0 and lag 1 residuals show an inverse relationship starting with row 5 (Fig. 11).

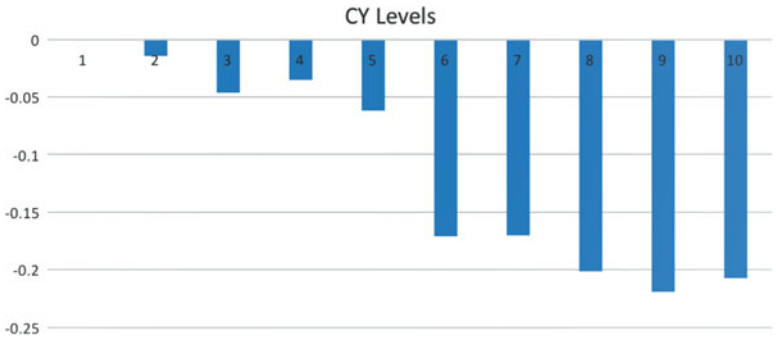


Fig. 10 Calendar year levels

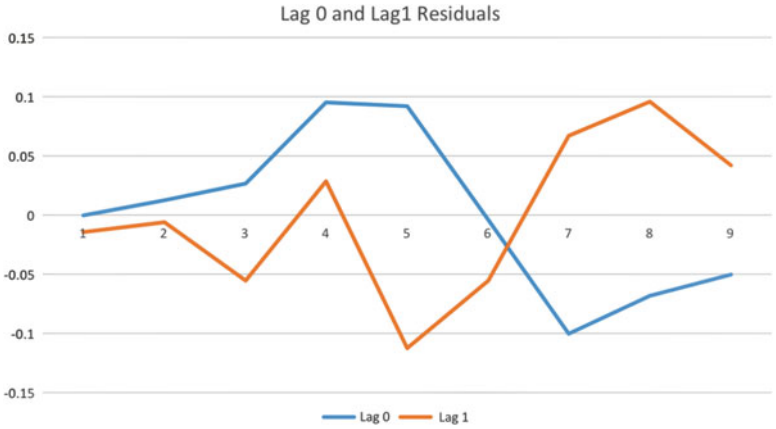


Fig. 11 Lag 0 and Lag 1 residuals by accident year

6.3 Model Extensions

Probably the most typical actuarial response to changes in payout pattern is to just use the data after the change. Meyers (2015) introduces modeling of changing payout patterns. With $y[n, u]$ = log of incremental claims for year n and lag u, one of his models can be written as:

$$y[n, u] = p[n] + q[u]z^{n-1} + \varepsilon_{n,u} \tag{5}$$

If $z = 1$, the payout pattern is constant, but if it is a bit above or below 1, the payout is speeding up or slowing down. This model does not include changes in trend, however, nor parameter reduction. One possible way to incorporate all of these effects is to add an interaction term between lag and accident year:

$$y[n, u] = p[n] + q[u] + w[n]x[u] + r[n + u] + \varepsilon_{n,u} \tag{6}$$

The slope changes for $u[n]$ and $x[u]$ in the interaction term were modeled as starting at the bottom and right, and built up going across the triangle right to left. The linear combination $q[u] + w[n]x[u]$ for the changing payout pattern is shown by cell.

The zeros at the bottom left are for identifiability and are the largest numbers in the triangle. A payout shift is seen from lag one, mostly to lag zero, but slightly to lag two as well. With the payout change modeled, the calendar-year levels below seem to be moving more uniformly. However, there is still a bigger change showing up in 1993 (Fig. 12).

At this point this model with interaction is still exploratory, but it does suggest such interactions may have a place in reserve triangle modeling (Fig. 13).

DY term after interaction									
-0.0846236	-0.2222793	-0.6875312	-0.8748535	-1.1997081	-1.594766	-1.7877335	-1.905103	-1.9601988	-2.0607415
-0.0547095	-0.2309704	-0.6827897	-0.8662319	-1.2002981	-1.5819962	-1.7773879	-1.8915242	-1.9601988	
-0.0562694	-0.2305172	-0.6830369	-0.8666815	-1.2002674	-1.5826621	-1.7779274	-1.8922323		
-0.0602586	-0.2293582	-0.6836692	-0.8678312	-1.2001887	-1.584365	-1.779307			
-0.0616277	-0.2289604	-0.6838862	-0.8682258	-1.2001617	-1.5849495				
-0.0420673	-0.2346434	-0.6807858	-0.8625882	-1.2005475					
-0.0177594	-0.2417056	-0.6769329	-0.8555824						
-0.0059299	-0.2451425	-0.6750579							
0	-0.2468654								
0									

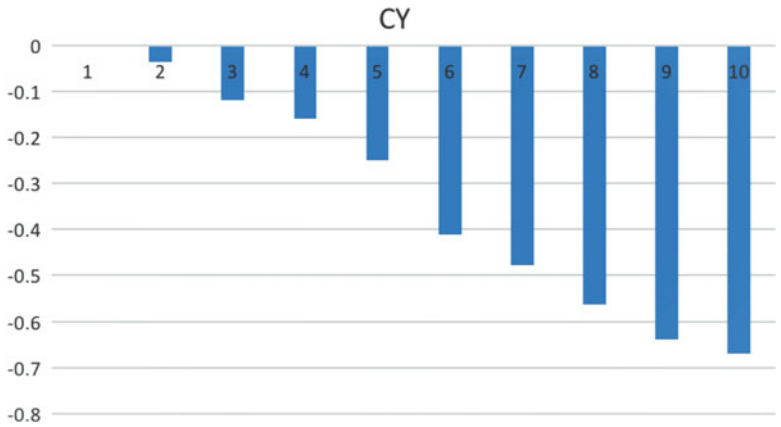


Fig. 12 Calendar year levels in model with changing payout patterns

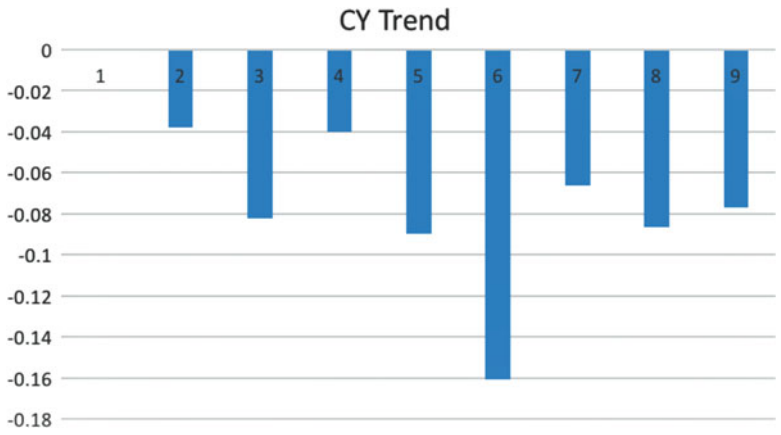


Fig. 13 Calendar year trends (slopes) in model with changing payout patterns

7 Conclusion

I like the maxim: All statements that begin with "All" are aimed at dramatic effect.

Still the idea that models may be only approximations but nonetheless can be useful is a key element of the shift towards pragmatism taking place in statistics. I am calling this the Robust Paradigm because of the notion that models need to be robust to effects that do not show up in the data at hand. This is broader than what usually is called robust statistics.

Assuming that the data is generated by the model process produces statistical tests that are mainly suggestive in this context. Out-of-sample testing is the requirement now. The availability of fast loo allows this to be standardized to a degree. Overfitting and so penalizing for too many parameters is no longer an issue when model performance out of sample is the focus.

But this is not traditional Bayesian either. Prior and posterior distributions are not statements of opinion. They are pieces of the story the model is telling us, and are as real as any other mathematical objects, such as quantum fields in the standard model of physics. And they are first and foremost pragmatic—helping to build a coherent narrative that provides insight into a process.

Parameter reduction now has classical and Bayesian modes. In the end the Bayesian approaches look more flexible and so more useful, particularly because of efficient loo.

The actuarial model with time variables is over-parameterized and so is a natural place for parameter reduction. This appears promising both for mortality and loss development applications. The more complex versions with interactions seem applicable to reserves, especially with payout pattern changes. Fairly extensive constraints are needed to get the parameters to do what they are meant to, however. There are a lot of possible overlaps and tradeoffs among parameters that need to be recognized explicitly if the models are going to perform as intended.

References

Lindstrom, M.J., Bates, D.M.: Nonlinear mixed effects models for repeated measures data. Biometrics **46**(3) (1990). ftp://www.biostat.wisc.edu/pub/lindstrom/papers/biometrics.1990.pdf

Meyers, G.: Stochastic loss reserving using Bayesian MCMC models. CAS Monograph Series, No. 1 (2015). http://www.casact.org/pubs/monographs/papers/01--Meyers.PDF

Osbourne, M.R., Presnell, B., Turlach, B.A.: On the lasso and its dual. J. Comput. Graph. Stat. **9**(2), 319–337 (2000). http://www.stat.washington.edu/courses/stat527/s13/readings/osborneetal00.pdf

Pereira, J.M., Basto, M., Ferreira da Silva, A.: The logistic lasso and ridge regression in predicting corporate failure. Proc. Econ. Financ. **39**, 634–641 (2016). http://www.sciencedirect.com/science/article/pii/S2212567116303100

Taylor, G., McGuire, G.: Stochastic loss reserving using generalized linear models. CAS Monograph Series, No. 3 (2016). http://www.casact.org/pubs/monographs/papers/03--Taylor.pdf

Ye, J.: On measuring and correcting the effects of data mining and model selection. J. Am. Stat. Assoc. **93**, 120–131 (1998)

Unlocking Reserve Assumptions Using Retrospective Analysis

Jeyaraj Vadiveloo, Gao Niu, Emiliano A. Valdez, and Guojun Gan

Abstract In this paper, we define a retrospective accumulated net asset random variable and mathematically demonstrate that its expectation is the retrospective reserve which in turn is equivalent to the prospective reserve. We further explore various properties of this retrospective accumulated net asset random variable. In particular, we find and demonstrate that this retrospective random variable can be used as a tool for helping us extract historical information on the pattern and significance of deviation of actual experience from that assumed for reserving purposes. This information can subsequently guide us as to whether it becomes necessary to adjust prospective reserves and the procedure to do so. The paper concludes, as an illustration, with a model of a block of in force policies with actual experience different from reserving assumptions and a suggested methodology on how prospective reserves could be adjusted based on the realized retrospective accumulated net asset random variable.

Keywords Life insurance reserves • Prospective loss • Retrospective accumulated net asset • Emerging mortality experience • Unlocking assumptions

1 Introduction

Reserves for life insurance products are funds set aside to meet the insurer's future financial obligations and they appear as a liability item on the insurer's balance sheet. This item usually represents a very large proportion of the insurance company's total liability and it is the task of the appointed actuary, responsible for the calculation of these reserves, to ensure that they are calculated according to well-accepted actuarial principles, within the guidelines set by the purpose of its calculation (e.g., statutory, tax), and that sufficient assets are available to back these reserves. See Atkinson and Dallas (2000, Chap. 6, pp. 313–356).

J. Vadiveloo • G. Niu • E.A. Valdez (✉) • G. Gan
University of Connecticut, Storrs, CT 06269-1009, USA
e-mail: jeyaraj.vadiveloo@uconn.edu; gao.niu@uconn.edu; emiliano.valdez@uconn.edu; guojun.gan@uconn.edu

J.A. Londoño et al. (eds.), *Actuarial Sciences and Quantitative Finance*, Springer Proceedings in Mathematics & Statistics 214, https://doi.org/10.1007/978-3-319-66536-8_2

Under old accounting rules, reserve basis and assumptions have typically been "locked-in" at policy issue so that they remain unchanged over time. However, it has become increasingly recognized that this "locked-in" principle can no longer be applicable under today's dynamic conditions. For example, under the Financial Accounting Standards (FAS) 97 and 120 for Generally Accepted Accounting Principles (GAAP), reserves can now be re-evaluated using what has been referred to as "dynamical unlocking" which allows for the replacement of original actuarial assumptions with a more realistic set of assumptions that accurately reflects historical experience when projecting for future years. See Financial Accounting Standards Board (1987).

The "locked-in" principle has also been historically applicable for statutory accounting, the basis that is used to value insurer's reserves and obligations to meet regulatory requirements for ensuring company solvency. Under old valuation standards, it has even been considered more deficient because the calculation of reserves has been static and formula-based. However, the National Association of Insurance Commissioners (NAIC), the organization responsible for formulating these uniform standards, has introduced in 2009 a new Standard Valuation Law (SVL) called Principle-Based Reserving (PBR). Under this PBR approach, insurance companies are now permitted to compute reserves by examining a wide range of more realistic future conditions, provided justified, and that the unlocking of reserve assumptions are permitted, again provided justified. This new valuation approach reflects the fact that insurance companies have been introducing more complex products to a more sophisticated market and that economic conditions are constantly evolving. See Manning (1990) and Mazyck (2013).

What these developments mean to the actuary is the need to continually evaluate historical experience and make necessary adjustments to the assumptions and reserves accordingly. The purpose of this article is to examine the use of a retrospective random variable to provide a guidance for unlocking reserve assumptions. For purposes of this article, we ignore the effect of expenses on reserves and focus on what has historically been called net level premiums reserves. Extension of concepts introduced in this article to reflect expenses should be straightforward, and our intent is to introduce first the concept so that it can be well explained more intuitively.

It is well known that net level premium reserves can be calculated prospectively and retrospectively at any duration for a policy that is in force. All major actuarial textbooks covering the mathematics of life contingencies demonstrate the equivalence between these two approaches based on an expected basis. See, for example, Bowers et al. (1986, Chap. 7, pp. 213–214) and Dickson et al. (2013, Chap. 7, pp. 220–225). To illustrate, consider a fully discrete n-year term insurance policy issued to a life aged x with a death benefit of M and an annual level premium of P determined according to the actuarial equivalence principle. At policy duration t, the prospective loss random variable is defined to be the difference between the present value of future benefits at time t (PVFB_t) and the present value of future premiums at time t (PVFP_t):

$$L_t^P = \text{PVFB}_t - \text{PVFP}_t, \tag{1}$$

where for our policy, we have

$$\text{PVFB}_t = M\, v^{K_{x+t}+1}\, I(K_{x+t} < n-t) \;\text{ and }\; \text{PVFP}_t = P\, \ddot{a}_{\overline{\min(K_{x+t}+1,\, n-t)}|},$$

where K_{x+t} refers to the curtate future lifetime of $(x+t)$ and $I(\cdot)$ is an indicator function. The expected value of this prospective loss random variable is the prospective reserve defined by

$$\text{E}(L_t^P) = \text{E}(\text{PVFB}_t) - \text{E}(\text{PVFP}_t) = M\, A^1_{x+t:\overline{n-t}|} - P\, \ddot{a}_{x+t:\overline{n-t}|} \tag{2}$$

and is referred to as the prospective net level premium reserve for this policy. Implicit in this formula is the assumption that the policyholder (x) has reached to survive t years. A straightforward algebraic manipulation of Eq. (2) leads us to the following equivalent expression of this reserve:

$$\text{Retrospective Reserve} = P\, \frac{\ddot{a}_{x:\overline{t}|}}{{}_tE_x} - M\, \frac{A^1_{x:\overline{t}|}}{{}_tE_x}, \tag{3}$$

where ${}_tE_x = v^t\, {}_tp_x$. Equation (3) is referred to as the retrospective net level premium reserve which gives the difference between the actuarial accumulated value of past premiums and the actuarial accumulated value of past benefits. Note that the mathematical equivalence of the retrospective and prospective reserve assumes that premiums at issue are determined based on the actuarial equivalence principle and that reserving assumptions equal pricing assumptions.

However, only the prospective reserve is defined as the expected value of a corresponding prospective loss random variable. Defining the corresponding retrospective accumulated net asset random variable that leads us to Eq. (3) has not appeared in the literature, and indeed, Dickson et al. (2013, Chap. 7, pp. 222–223) and Gerber (1976) recognize the difficulty of defining such a random variable. In this paper, we define a retrospective accumulated net asset random variable whose expectation leads us to the retrospective reserve and is therefore equal to the prospective reserve. We are also able to intuitively provide an interpretation to this loss random variable. We further explore various properties of the retrospective accumulated net asset random variable and how its realized value provides valuable information on how prospective reserves may be established.

In this paper, we develop a formal definition of a retrospective accumulated net asset random variable whose expected value is equal to the retrospective reserve, which in turn equals the prospective reserve. However, while both the accumulated net asset random variable and prospective loss random variable have equal expectations, the probability distributions of both random variables are entirely different. The paper will provide an intuitive explanation and additional insight as to what the retrospective accumulated net asset random variable is measuring and how its distribution differs from the prospective loss random variable over time. More importantly, the paper additionally explores how the retrospective accumulated net

asset random variable could provide information on a company's historical claim experience and how the prospective reserve at any duration t should be adjusted if actual experience over the past t years differs from reserving assumptions. The retrospective accumulated net asset random variable as defined in this paper can help an insurance company in developing a claims tracking and monitoring process and provide a systematic procedure of adjusting future reserves to reflect actual experience. This procedure can then be implemented to meet valuation standards according to Principle-Based Reserving.

This paper has been structured as follows. Section 2 develops the theoretical foundation for defining the retrospective accumulated net asset random variable. Here, we demonstrate how this definition differs from the more familiar prospective loss random variable, though we also show that the two are always equal in expectation. This equality in expectation hinges on the premium being determined according to the actuarial equivalence principle. Section 3 extends the discussion of the retrospective accumulated net asset random variable in the case where we have a portfolio of insurance policies. This further gives us a natural interpretation of the retrospective accumulated net asset random variable. Furthermore, in this section, we show how one can derive the mean and variance of the retrospective accumulated net asset random variable for a portfolio that may vary in the amounts of death benefits and issue ages. This is important because we demonstrate how the standard deviation of the retrospective may be used to unlock the assumption of mortality so that prospective reserves may be adjusted accordingly. The adjustment in our demonstration may be arbitrary, for the moment, but it allows us to systematically make the adjustment. We conclude in Sect. 4.

2 Formulation

2.1 Defining the Retrospective Accumulated Net Asset Random Variable

The retrospective accumulated net asset random variable is best understood with a simple illustration. Extension to the case of other forms of insurance will be rather straightforward and we will examine a few of these other cases.

Consider a fully discrete n-year term insurance policy issued to a life aged x with a death benefit of M and an annual level premium of P determined according to the actuarial equivalence principle. For those unfamiliar with the concept of fully discrete, this refers to the death benefit being paid at the end of the year of death and that level premiums are paid at the beginning of each year the policyholder is alive. See Bowers et al. (1986, Chap. 7, pp. 215–221) and Gerber (1997, Chap. 6, pp. 59–73).

For a policyholder age x, denote his curtate future lifetime random variable by K_x. For $K_x < t$, the policyholder dies before reaching age $x + t$ and in this case, we define the retrospective accumulated net asset random variable to be

$$L_t^R = \frac{1}{{}_tp_x}\left[P\ddot{a}_{\overline{K_x+1}|}(1+i)^t - M(1+i)^{t-K_x-1}\right], \tag{4}$$

where ${}_tp_x$ is the probability that policyholder (x) survives for t years. The first term $P\ddot{a}_{\overline{K_x+1}|}(1+i)^t$ clearly refers to the accumulated value at time t of all past premiums paid before death while the second term $B(1+i)^{t-K_x-1}$ refers to the accumulated value of the death benefit, paid at the end of the year of death, at time t.

In the case where $K_x \geq t$, we define the retrospective accumulated net asset random variable to be simply a constant equal to

$$L_t^R = \frac{P\ddot{a}_{\overline{t}|}(1+i)^t}{{}_tp_x}. \tag{5}$$

We can express this retrospective accumulated net asset random variable more succinctly as

$$\begin{aligned} L_t^R &= \frac{1}{{}_tp_x}\left[P(1+i)^t\left(\ddot{a}_{\overline{K_x+1}|}\cdot I(K_x<t) - \ddot{a}_{\overline{t}|}\cdot I(K_x \geq t)\right) - M(1+i)^{t-K_x-1}\cdot I(K_x<t)\right] \\ &= \frac{1}{{}_tp_x}\left[P(1+i)^t\,\ddot{a}_{\overline{\min(K_x+1,t)}|} - M(1+i)^{t-K_x-1}\cdot I(K_x<t)\right] \end{aligned} \tag{6}$$

In the case where $K_x \geq n$, the policyholder would have survived the term of the policy and in which case, L_t^R would still be Eq. (5).

It is therefore straightforward to interpret the retrospective accumulated net asset random variable. In this case, it can be viewed as the share per survivor of the `accumulated net assets` per \$1 of insurance at duration t. A similar concept of an expected share per survivor within the context of group benefits has been considered in Ramsay (1993) and Arias Lopez and Garrido (2001). In contrast, the prospective loss random variable can be viewed as the share per survivor of the `present value of net liabilities` per \$1 of insurance at duration t. We will define the expectation of this retrospective accumulated net asset random variable, $\mathrm{E}(L_t^R)$, as the retrospective reserve.

Using formulas from mathematics of life contingencies, it is straightforward to prove the equivalence between prospective and the retrospective reserve. Note that we can express Eq. (6) as

$$L_t^R = \frac{1}{v^t\,{}_tp_x}\left[P\,\ddot{a}_{\overline{\min(K_x+1,t)}|} - Mv^{K_x+1}\cdot I(K_x<t)\right] \tag{7}$$

so that we write

$$\begin{aligned}\mathrm{E}(L_t^R) &= \frac{1}{v^t\,{}_tp_x}\left\{P\,\mathrm{E}\left[\ddot{a}_{\overline{\min(K_x+1,t)|}}\right] - M\mathrm{E}\left[v^{K_x+1}\cdot I(K_x<t)\right]\right\}\\ &= \frac{1}{{}_tE_x}\left(P\,\ddot{a}_{x:\overline{t}|} - MA^{\,1}_{x:\overline{t}|}\right).\end{aligned}$$

According to the actuarial equivalence principle, we have $P\,\ddot{a}_{x:\overline{n}|} = MA^{\,1}_{x:\overline{n}|}$. It follows therefore that

$$\begin{aligned}\mathrm{E}(L_t^R) &= \frac{1}{{}_tE_x}\left(P\,\ddot{a}_{x:\overline{t}|} - MA^{\,1}_{x:\overline{t}|} - P\,\ddot{a}_{x:\overline{n}|} + MA^{\,1}_{x:\overline{n}|}\right)\\ &= \frac{1}{{}_tE_x}\left[M\left(A^{\,1}_{x:\overline{n}|} - A^{\,1}_{x:\overline{t}|}\right) - P\left(\ddot{a}_{x:\overline{n}|} - \ddot{a}_{x:\overline{t}|}\right)\right]\\ &= M\,A^{\,1}_{x+t:\overline{n-t}|} - P\,\ddot{a}_{x+t:\overline{n-t}|} = \mathrm{E}(L_t^P).\end{aligned}$$

Notice that although the expectations are equal at any duration t, the probability distributions of the two random variables are not. Indeed at policy issue, that is, at $t = 0$, it is easy to see that $L_0^R = 0$ although

$$L_0^P = B\,v^{K_x+1}\,I(K_x<n) - P\,\ddot{a}_{\overline{\min(K_x+1,n)|}}$$

and is not necessarily always equal to zero. However, by the equivalence principle, it follows directly that $\mathrm{E}(L_0^P) = 0$. Because at policy issue there should be no net assets accumulated, we easily see that $L_0^R = 0$. Indeed, this alone shows that the two random variables are different in distribution.

In contrast, we see that at policy maturity $t = n$, the prospective loss is $L_n^P = 0$ since there is no more future net liabilities. However, the retrospective accumulated net asset random variable at policy maturity is

$$L_n^R = \frac{1}{{}_nE_x}\left[P\,\ddot{a}_{\overline{\min(K_x+1,n)|}} - B\,v^{K_x+1}\,I(K_x<n)\right]$$

which also is not necessarily equal to zero although it has zero expectation again because of the equivalence principle.

2.2 Understanding Differences Between the Prospective Loss and the Retrospective Accumulated Net Asset

To further understand the difference between these two random variables, consider a fully discrete 25-year term insurance policy issued to age $x = 40$ and assume mortality follows the Gompertz law with

$$\mu_{40+t} = B \cdot c^{40+t}, \quad \text{for } t \geq 0,$$

where $B = 0.0000429$ and $c = 1.1070839$. We examine the differences between the prospective loss and retrospective accumulated net asset random variables at the end of year 10. For illustration purpose, we assume that the annual effective interest rate is 5% and the death benefit, payable at the end of the year of death, is \$100,000.

First, note that the prospective random variable is based on the future lifetime of the policyholder from duration t. This refers to the loss that is conditional on survival of the policyholder at time t and we are looking at the difference between the present value of future benefits yet to be paid and future premiums yet to be collected. In contrast, the retrospective accumulated net asset random variable is based on the future lifetime of the policyholder from issue and this is because we must look back at what happened to the difference in the accumulation of premiums and benefits paid in the past prior to duration t. This explains why, as earlier stated, the prospective loss random variable can be viewed as the share per survivor of the `present value of net liabilities` per \$1 of insurance at duration t while the retrospective accumulated net asset random variable as the share per survivor of the `accumulated net assets` per \$1 of insurance at duration t.

We can further visualize this difference with the help of Fig. 1 where we compare the realized prospective loss and retrospective accumulated net asset at duration t given the policyholder dies at a point in time. For the prospective loss, because the random variable is conditional on survival at time t, we consider death at each year after reaching age $x + t$. For the retrospective accumulated net asset random variable, we consider death at each year after issue age x but up to age $x + t$. Despite this difference in the future lifetime random variables, we see that earlier deaths for the prospective case generates larger positive net liabilities than later deaths

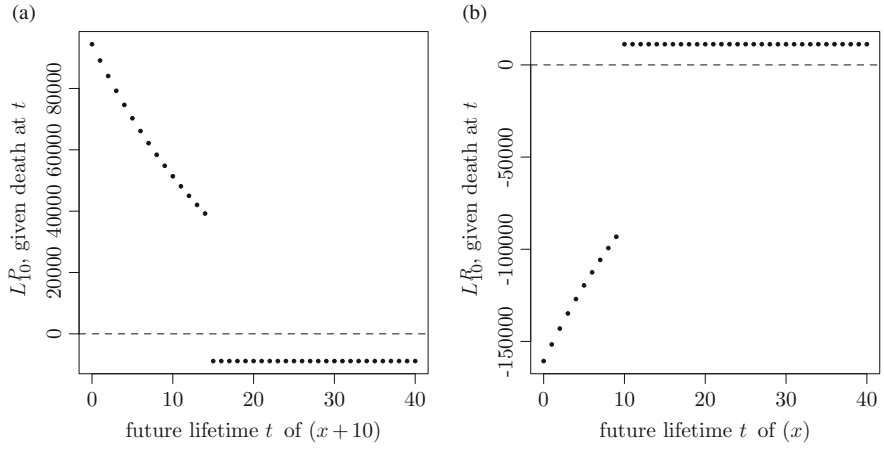


Fig. 1 Comparison of realized prospective loss and retrospective accumulated net asset at duration 10. (**a**) Prospective. (**b**) Retrospective

and this pattern is quite apparent in our example. For the retrospective case, earlier deaths generate fewer accumulated assets than later deaths. This can be intuitively explained by the fact that for early deaths, collected premiums will be fewer and that the death benefit is accumulated for a longer period from death to the duration in consideration; in this case, the duration is 10 years.

It is also interesting to note that for the prospective case, the random variable is constant after the term of the policy. This is because the prospective loss will have simply consisted of the present value, at duration 10 years, of future premiums collected up to the term of the policy since the death benefit portion will have always been zero. In contrast for the retrospective case, the random variable is constant for deaths after duration 10. This is because the retrospective accumulated net asset will have simply consisted of the share of the survivors of the accumulated value, at duration 10, of all premiums collected from issue till duration 10.

Finally, it is well worth examining the comparison between the shape of the distributions between the prospective loss and retrospective accumulated net asset. In Fig. 2, using the same set of assumptions to develop Fig. 1 and the Monte Carlo simulation, we compare the histograms between these two loss random variables. Observe the noticeably high proportion of a negative net liability in the prospective case and the noticeably high proportion of a positive net asset accumulation in the retrospective case. In the prospective case, this negative net liability is attributable to those survivors by the end of the policy term and beyond. In the retrospective case, this positive net asset accumulation is attributable to those survivors at duration 10 and beyond.

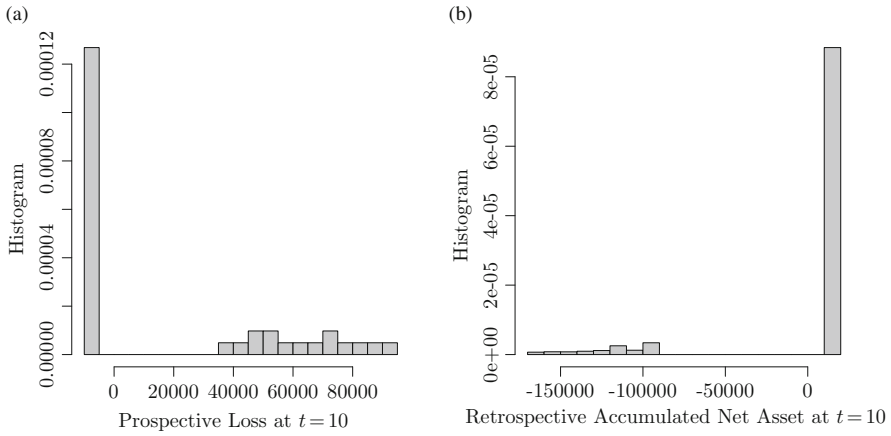


Fig. 2 Distribution of prospective loss and retrospective accumulated net asset at duration 10. (**a**) Prospective. (**b**) Retrospective

2.3 Numerical Illustration

To even further understand the retrospective accumulated net asset random variable, we consider a numerical illustration. For this purpose, we consider a fully discrete 20-year term insurance policy issued to age $x = 45$ with a death benefit of $M = \$1000$. For mortality assumption, we consider a table widely used in the industry for valuation purposes: the 2015 VBT Unismoke Age Nearest Birthday (ANB) mortality table. With interest rate equal to $i = 5\%$, we find that, using the equivalence principle, the net annual premium $P = 2.58$ per \$1000 of insurance.

Table 1 below shows the distribution of the retrospective accumulated net asset random variable at time $t = 10$ for the 11 possible realizations of the retrospective accumulated net asset random variable, L_{10}^R, for durations $1, 2, \ldots 10, 11$ and later. According to this calculation, we find that

$$\mathrm{E}[L_{10}^R] = 17.19 \text{ and } \mathrm{SD}[L_{10}^R] = 145.42$$

per \$1000 of insurance.

Table 2 shows the mean and standard deviation of both the retrospective and prospective loss random variables per \$1000 of insurance for the durations $t = 1, 2, \ldots 20$. Since the prospective loss random variable is a well-known random variable in the actuarial literature, we will assume the reader is familiar with its distribution for the simple insurance example we have illustrated. This table also demonstrates that for a given duration t, we can see that the expectations of the prospective loss and retrospective accumulated net asset are equal. However, the standard deviations for the same duration are not necessarily the same. In general, the standard deviation of the retrospective accumulated net asset random variable is smaller than the standard deviation of the prospective loss random variable in the early durations but it reverses in the later durations. Also, the standard deviation of the retrospective accumulated net asset random variable steadily increases as

Table 1 Distribution of the retrospective accumulated net asset random variable per \$1000 at duration 10, where $x = 45$, $n = 20$, $i = 5\%$, and gender = male

Duration	Retrospective accumulated net asset	
t	L_t^R	Probability
0	−1,569.77	0.0005
1	−1,490.75	0.0007
2	−1,415.49	0.0009
3	−1,343.82	0.0011
4	−1,275.56	0.0013
5	−1,210.55	0.0015
6	−1,148.63	0.0017
7	−1,089.66	0.0020
8	−1,033.51	0.0022
9	−980.02	0.0025
≥ 10	34.62	0.9856

Table 2 Mean and standard deviation of retrospective accumulated net asset and prospective loss random variables per $1000, where $x = 45$, $n = 20$, $i = 5\%$, and gender = male

Duration	Retrospective accumulated net asset RV		Prospective loss RV	
t	Mean	Standard deviation	Mean	Standard deviation
1	2.24	21.68	2.24	138.35
2	4.37	34.96	4.37	142.91
3	6.39	47.71	6.39	147.07
4	8.31	60.38	8.31	150.90
5	10.16	73.05	10.16	154.51
6	11.91	86.09	11.91	157.81
7	13.51	99.79	13.51	160.69
8	14.94	114.21	14.94	163.08
9	16.18	129.36	16.18	164.95
10	17.19	145.42	17.19	166.14
11	17.92	162.43	17.92	166.53
12	18.30	180.59	18.30	165.85
13	18.26	200.01	18.26	163.81
14	17.76	220.68	17.76	160.18
15	16.71	242.77	16.71	154.42
16	14.95	266.51	14.95	145.65
17	12.45	291.92	12.45	132.81
18	9.18	318.95	9.18	114.15
19	5.06	347.68	5.06	85.00
20	0.00	378.27	0.00	0.00

duration increases, but this is not the case for the prospective loss random variable. Such pattern is to be expected as we have also demonstrated in our comparison in the previous section.

2.4 Extensions to Other Forms of Insurance

First, consider the case of a fully discrete whole life insurance policy. One can easily show the extension is straightforward because one can simply think of this as a term insurance with an infinite maturity. Premiums continue to be collected until death and policy expires at the end of the year of death of the policyholder.

In this case, we can express the retrospective accumulated net asset random variable in a similar fashion to Eq. (6). The only difference has to do with the value of the net annual premium. Using the equivalence principle, this leads us to

$$P\ddot{a}_x = MA_x \tag{8}$$

To demonstrate that the expectation of this retrospective accumulated net asset random variable is equal to that of the prospective loss random variable, we follow the same procedure as in the fully discrete term insurance.

$$\begin{aligned}\mathrm{E}(L_t^R) &= \frac{1}{{}_tE_x}\left(P\,\ddot{a}_{x:\overline{t}|} - MA^{1}_{x:\overline{t}|} - P\,\ddot{a}_x + MA_x\right)\\ &= M\,A_{x+t} - P\,\ddot{a}_{x+t} = \mathrm{E}(L_t^P).\end{aligned}$$

In the case of a fully continuous whole life insurance, one can also easily develop the retrospective accumulated net asset random variable at duration t by defining it to be

$$L_t^R = \frac{1}{{}_tp_x}\left[\overline{P}\,(1+i)^t\,\overline{a}_{\overline{\min(T_x,t)}|} - M(1+i)^{t-T_x}\cdot I(T_x < t)\right] \tag{9}$$

where $\overline{P}$ denotes the annual premium rate and T_x is the future lifetime of (x). The corresponding prospective loss random variable in this case is defined to be

$$L_t^P = M\,v^{T_{x+t}} - \overline{P}\,\overline{a}_{\overline{T_{x+t}}|} \tag{10}$$

where T_{x+t} is the future lifetime of $(x+t)$.

Analogous to the development of the fully discrete, we have the retrospective reserve, equal to the expectation of the retrospective accumulated net asset random variable, for a fully continuous whole life as follows

$$\mathrm{E}(L_t^R) = \overline{P}\,\frac{\overline{a}_{x:\overline{t}|}}{{}_tE_x} - M\,\frac{\overline{A}^{1}_{x:\overline{t}|}}{{}_tE_x}, \tag{11}$$

and the prospective reserve, equal to the expected value of the prospective loss random variable, is

$$\mathrm{E}(L_t^P) = M\,\overline{A}_{x+t} - \overline{P}\,\overline{a}_{x+t}. \tag{12}$$

According to the actuarial equivalence principle, we have $\overline{P}\,\overline{a}_x = M\overline{A}_x$. Following similar proof as in the fully discrete case, it is straightforward to show the two expectations are equal.

To close this section, it is interesting to consider the case of an n year pure endowment policy where a benefit of 1 is payable at maturity if the policyholder, age x, survives then. Here we assume that premiums are payable annually at the rate of $\overline{P}$ and are determined according to the actuarial equivalence principle so that we have

$$\overline{P} = \frac{{}_nE_x}{\overline{a}_{x:\overline{n}|}}.$$

In this case, we write the retrospective accumulated net asset random variable at time $t < n$ as

$$L_t^R = \overline{P}\frac{1}{{}_tp_x}\overline{a}_{\overline{T_x}|}(1+i)^t,$$

for $T_x < t$ and

$$L_t^R = \overline{P}\frac{1}{{}_tp_x}\overline{a}_{\overline{t}|}(1+i)^t,$$

for $T_x \geq t$.

As clearly interpreted in this paper, this random refers to the "the share per survivor of the `accumulated net assets` per \$1 of insurance at duration t". For those people who died before duration t, they would have paid total premiums up to their time of death. For those who have survived to duration t, they would have paid total premiums up to time t. In either case, no pure endowment benefit has yet been paid since $t < n$. Hence, the interpretation as stated. This same random variable can be succinctly written as

$$L_t^R = \overline{P}\frac{1}{{}_tE_x}\overline{a}_{\overline{\min(T_x,t)}|}. \tag{13}$$

3 Reserve Adjustment Based on the Retrospective Accumulated Net Asset Random Variable for a Portfolio

Consider a portfolio of m independent policies all issued with possible varying death benefit amounts and issue ages. Denote the benefit amount, typically called face amount in practice, for the ith policy byM_i and the aggregate retrospective accumulated net asset variable at duration t for this portfolio by $L_{\text{agg},t}^R$. It is not difficult to see that if $L_{i,t}^R$ is the retrospective accumulated net asset random variable per dollar of death benefit, then the ith policy retrospective accumulated net asset random variable can be expressed as $M_i \times L_{i,t}^R$ so that the aggregate retrospective accumulated net asset random variable for the portfolio can be expressed as

$$L_{\text{agg},t}^R = \sum_{i=1}^{m} M_i \times L_{i,t}^R$$

Dividing this by the total face amount of $\sum_{i=1}^{m} M_i$, we get the aggregate retrospective accumulated net asset per dollar of insurance:

$$L^R_{\text{agg},1,t} = \frac{L^R_{\text{agg},t}}{\sum_{i=1}^m M_i} = \sum_{i=1}^m \frac{M_i}{\sum_{i=1}^m M_i} \times L^R_{i,t} = \sum_{i=1}^m p_i \times L^R_{i,t},$$

where

$$p_i = \frac{M_i}{\sum_{i=1}^m M_i} \text{ for } i = 1, 2 \ldots m.$$

Assuming independent future lifetimes of all individual policyholders within the portfolio, then aggregate mean per dollar of insurance is

$$\mathrm{E}(L^R_{\text{agg},1,t}) = \sum_{i=1}^m p_i \times \mathrm{E}(L^R_{i,t}) \tag{14}$$

and aggregate variance per squared dollar of insurance is

$$\mathrm{Var}(L^R_{\text{agg},1,t}) = \sum_{i=1}^m p_i^2 \times \mathrm{Var}(L^R_{i,t}). \tag{15}$$

These results simply demonstrate that the mean and the standard deviation of the retrospective accumulated net asset random variable per dollar of insurance of any portfolio of policies that were issued in the same year, can be analytically determined from the mean and standard deviation of the retrospective accumulated net asset random variable per dollar of insurance of the individual policies. These results have been heavily applied in the illustration of our portfolio development and reserve adjustment in the subsequent subsections.

3.1 Interpretation of the Retrospective Accumulated Net Asset Random Variable

The retrospective accumulated net asset random variable can be best interpreted by modeling a portfolio of policies with the same issue age x. Assuming that the only decrement is death, then at duration t, there are two values that could be generated from the model:

(a) accumulated net assets (i.e. accumulated premiums less accumulated death benefits) at $x + t$ based on the actual mortality experience of the portfolio in the first t durations, and
(b) expected number of policies remaining in force in duration t.

Then the realized retrospective accumulated net asset random variable is the ratio of (a) to (b) above, and it represents the share per survivor of the realized net assets at duration t. The distribution of the retrospective accumulated net asset random variable can be obtained by generating all possible realizations of this ratio (a)/(b). It is apparent that this cannot be done analytically, but the distribution of the retrospective accumulated net asset random variable can be obtained via simulation.

Table 3 shows the mean, standard deviation and various quantiles of interest of the retrospective accumulated net asset random variable per $1000 of face amount at various durations for a portfolio of 100 term insurance policies at each duration, issued at age 45 for face amount $100,000. For this purpose, we generated mortality patterns according to the 2015 VBT Unismoke Age Nearest Birthday (ANB) mortality table. The quantiles we are showing in Table 4 are mean $\pm$ 0.1*SD, mean $\pm$ 0.2*SD, mean $\pm$ 0.5*SD, mean $\pm$ SD and mean $\pm$ 3*SD, where SD refers to the standard deviation.

Figure 3 provides an interesting visualization of how the mean and standard deviation of the retrospective accumulated net asset random variable emerge over a period of duration 20. A few observations can be made here. First, for a term insurance policy, the retrospective reserve starts small and follows a parabolic pattern. At maturity, the retrospective reserve is equal to zero. Finally, it is interesting to note that standard deviation increases with duration, thus the widening of the confidence band. This increase with duration can be explained by the fact that we become increasingly uncertain of the retrospective accumulated net asset for later durations. In this article, we suggest to use such confidence bands to make the necessary adjustment to prospective reserves. This increasing standard deviation over time implies that as we accumulate enough experience over time, enough information will become available to give us greater confidence of making the necessary adjustment.

Table 4 shows the same results for the prospective loss random variable per $1000 of face amount by analyzing the future present value of net liabilities per policy at duration t based on 100 in force policies at duration t that were issued t years ago with all policies at issue age 45.

In comparing Tables 3 and 4, we can make the following inferences:

- The retrospective accumulated net asset random variable always satisfies the condition that

$$\text{E(retrospective accumulated net asset random variable)} = \text{E(prospective loss random variable)}$$

- Since all policies have the same face amount, the retrospective (and prospective) reserve per $1000 is equal to the reserve for a single $1000 face amount policy. However, the standard deviation per $1000 equals the corresponding SD for a single $1000 face amount policy divided by the square root of the number of policies in the portfolio (i.e., 10 in this example). This conforms to our earlier results on how the mean and standard deviation of the retrospective accumulated

Table 3 Mean, standard deviation and quantiles of the retrospective accumulated net asset random variable per $1000 for a portfolio, where $x = 45$, $n = 20$, $M = \$100,000$, $i = 5\%$, gender = male, and number of policies = 100

Duration	Mean	SD	Mean −0.1*SD	Mean +0.1*SD	Mean −0.2*SD	Mean +0.2*SD	Mean −0.5*SD	Mean +0.5*SD	Mean −SD	Mean +SD	Mean −3*SD	Mean +3*SD
1	2.24	2.17	2.03	2.46	1.81	2.68	1.16	3.33	0.08	4.41	−4.49	8.75
2	4.37	3.50	4.02	4.72	3.67	5.07	2.62	6.12	0.88	7.87	−8.74	14.86
3	6.39	4.77	5.91	6.87	5.43	7.34	4.00	8.77	1.62	11.16	−12.78	20.70
4	8.31	6.04	7.71	8.91	7.10	9.52	5.29	11.33	2.27	14.35	−16.62	26.42
5	10.16	7.30	9.43	10.89	8.70	11.62	6.51	13.81	2.86	17.47	−20.32	32.07
6	11.91	8.61	11.05	12.77	10.19	13.63	7.60	16.21	3.30	20.52	−23.82	37.74
7	13.51	9.98	12.51	14.51	11.51	15.51	8.52	18.50	3.53	23.49	−27.02	43.45
8	14.94	11.42	13.80	16.08	12.65	17.22	9.23	20.65	3.52	26.36	−29.87	49.20
9	16.18	12.94	14.89	17.48	13.60	18.77	9.71	22.65	3.25	29.12	−32.37	54.99
10	17.19	14.54	15.73	18.64	14.28	20.10	9.92	24.46	2.65	31.73	−34.38	60.81
11	17.92	16.24	16.30	19.55	14.67	21.17	9.80	26.04	1.68	34.17	−35.84	66.65
12	18.30	18.06	16.50	20.11	14.69	21.91	9.27	27.33	0.24	36.36	−36.60	72.48
13	18.26	20.00	16.26	20.26	14.26	22.26	8.26	28.26	−1.74	38.26	−36.51	78.26
14	17.76	22.07	15.55	19.96	13.34	22.17	6.72	28.79	−4.31	39.83	−35.51	83.96
15	16.71	24.28	14.28	19.13	11.85	21.56	4.57	28.84	−7.57	40.98	−33.41	89.54
16	14.95	26.65	12.29	17.62	9.62	20.28	1.63	28.28	−11.70	41.61	−29.91	94.91
17	12.45	29.19	9.53	15.37	6.61	18.29	−2.15	27.05	−16.74	41.64	−24.90	100.03
18	9.18	31.89	5.99	12.37	2.80	15.55	−6.77	25.12	−22.72	41.07	−18.35	104.86
19	5.06	34.77	1.59	8.54	−1.89	12.02	−12.32	22.45	−29.70	39.83	−10.13	109.37
20	0.00	37.83	−3.78	3.78	−7.57	7.57	−18.91	18.91	−37.83	37.83	0.00	113.48

Table 4 Mean, standard deviation and quantiles of prospective loss random variable per $1000 for a portfolio, where $x = 45$, $n = 20$, $M = \$100,000$, $i = 5\%$, gender = male, and number of policies = 100

Duration	Mean	SD	Mean −0.1*SD	Mean +0.1*SD	Mean −0.2*SD	Mean +0.2*SD	Mean −0.5*SD	Mean +0.5*SD	Mean −SD	Mean +SD	Mean −3*SD	Mean +3*SD
1	2.24	13.83	0.86	3.63	−0.52	5.01	−4.67	9.16	−11.59	16.08	−4.49	43.75
2	4.37	14.29	2.94	5.80	1.51	7.23	−2.77	11.52	−9.92	18.66	−8.74	47.24
3	6.39	14.71	4.92	7.86	3.45	9.33	−0.97	13.74	−8.32	21.09	−12.78	50.51
4	8.31	15.09	6.80	9.82	5.29	11.33	0.76	15.85	−6.78	23.40	−16.62	53.58
5	10.16	15.45	8.62	11.71	7.07	13.25	2.43	17.89	−5.29	25.61	−20.32	56.51
6	11.91	15.78	10.33	13.49	8.75	15.06	4.02	19.80	−3.87	27.69	−23.82	59.25
7	13.51	16.07	11.90	15.12	10.30	16.72	5.48	21.54	−2.56	29.58	−27.02	61.72
8	14.94	16.31	13.31	16.57	11.68	18.20	6.78	23.09	−1.37	31.25	−29.87	63.86
9	16.18	16.50	14.53	17.83	12.88	19.48	7.94	24.43	−0.31	32.68	−32.37	65.67
10	17.19	16.61	15.53	18.85	13.87	20.51	8.88	25.50	0.57	33.80	−34.38	67.03
11	17.92	16.65	16.26	19.59	14.59	21.25	9.60	26.25	1.27	34.58	−35.84	67.88
12	18.30	16.58	16.64	19.96	14.98	21.62	10.01	26.59	1.72	34.89	−36.60	68.05
13	18.26	16.38	16.62	19.90	14.98	21.53	10.07	26.45	1.88	34.64	−36.51	67.40
14	17.76	16.02	16.16	19.36	14.55	20.96	9.75	25.77	1.74	33.77	−35.51	65.81
15	16.71	15.44	15.16	18.25	13.62	19.79	8.99	24.43	1.26	32.15	−33.41	63.03
16	14.95	14.56	13.50	16.41	12.04	17.87	7.67	22.24	0.39	29.52	−29.91	58.65
17	12.45	13.28	11.12	13.78	9.79	15.11	5.81	19.09	−0.83	25.73	−24.90	52.29
18	9.18	11.41	8.03	10.32	6.89	11.46	3.47	14.88	−2.24	20.59	−18.35	43.42
19	5.06	8.50	4.21	5.91	3.36	6.76	0.81	9.31	−3.44	13.56	−10.13	30.56
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

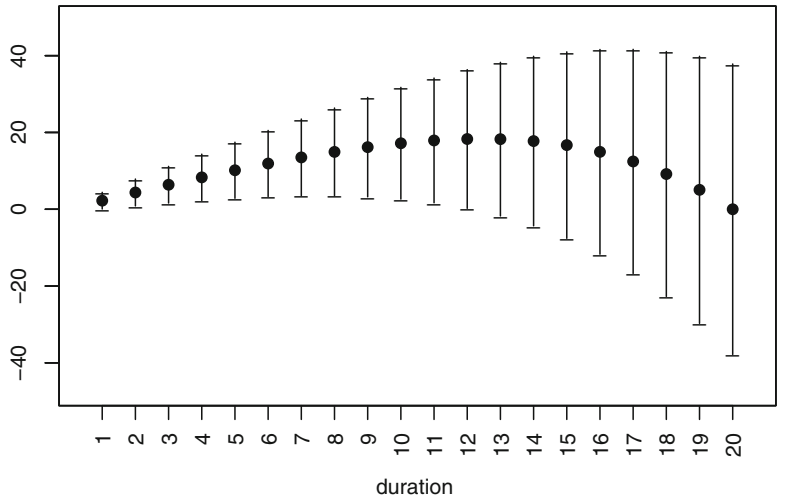


Fig. 3 Mean ± one standard deviation of the retrospective accumulated net asset random variable

net asset random variable for a portfolio of policies may be conveniently calculated.

- There are variations in the standard deviations of the retrospective accumulated net asset and prospective loss random variables by duration.
- There are variations in the quantiles of the retrospective and prospective loss random variables by duration.

This leads us to the next couple of questions. Based on how we have defined the retrospective accumulated net asset random variable, what does it really mean from an insurance company's perspective? Furthermore, what can we learn from the volatility of the retrospective accumulated net asset random variable in setting the prospective reserves from an insurer's perspective.

3.2 Implications of the Retrospective Accumulated Net Asset Random Variable for Insurers

The retrospective reserve in the actuarial literature has been viewed as algebraically equivalent to the prospective reserve in expectation and a convenient alternative to determining policy reserves for certain product designs. By creating a retrospective accumulated net asset random variable, we hope to help increase the importance of the retrospective reserve as the mean of the distribution of the accumulated net assets per $1000 of insurance. This is a useful random variable for insurers to analyze in evaluating historical claims experience and determining how to set, or reset, prospective reserves.

Specifically, if the realized retrospective accumulated net asset random variable lies outside some pre-established confidence band for the retrospective accumulated net asset random variable, then the prospective reserve could be adjusted to reflect the fact that actual historical experience is significantly different from reserving assumptions. This could become the regulatory basis for adjusting future reserves in accordance with Principles Based Reserving. This can also form the basis of a claims tracking and monitoring process for an insurer.

In the illustration that follows, we consider a portfolio of term life insurance policies. For purpose of setting up the mortality pattern, we consider the same valuation table we have previously used: the 2015 VBT Unismoke Age Nearest Birthday (ANB) mortality table.

In order for an insurance company to implement a process by which prospective reserves are adjusted for an in force block of policies in a systematic manner to reflect the realized retrospective accumulated net asset random variable, the following steps have to be done:

1. The in force block has to be broken down into issue year groupings and by plan of insurance.
2. For a given issue year and plan of insurance, the historical premiums and death claims paid have to be accumulated to the valuation date to determine the realized retrospective reserve per $1000 of face amount.
3. The realized retrospective reserve determined in Step (2) above will have to be compared to the retrospective accumulated net asset random variable per $1000 of face amount confidence band at a pre-established level of confidence (e.g., mean $\pm$ SD). Note that both the mean and standard deviation of the confidence band vary by policy duration and we can use the result to determine the portfolio confidence band.
4. To recognize the fact that as duration from issue date to valuation date increases, the retrospective accumulated net asset random variable is based on more credible historical experience, the confidence bands could vary by duration. For example, the later durations (i.e., earlier issues) could use a tighter confidence band while earlier durations (i.e., later issues) could use a wider confidence band.
5. A possible (and certainly hypothetical) rule for adjusting the prospective reserves for this issue year block and plan of insurance could be as follows:

    - If the realized retrospective accumulated net asset random variable falls within the pre-established confidence band around the mean, then no adjustment is made to the prospective reserve.
    - If the realized retrospective accumulated net asset random variable exceeds the upper confidence band by $1 per $1000 of insurance, then the prospective reserve for the issue year block can be reduced by $1 per $1000 of insurance.
    - If the realized retrospective accumulated net asset random variable is below the lower confidence band by $1 per $1000 of insurance, then the prospective reserve for the issue year block should be increased by $1 per $1000 of insurance.

An area of further research that has not been explored in detail in this paper is developing a more systematic process of determining the width of the confidence interval (CI) by duration for the retrospective accumulated net asset random variable. One possible approach is to make some type of a credibility adjustment similar in concept to credibility concepts of adjusting expected claims based on past claims experience. There is certainly additional research work needed in this area. See, for example, a method used for variable annuity products in Longley-Cook et al. (2001). However, here we offer some possible approaches:

1. An overall consistency requirement is that the later the policy duration, the tighter the confidence interval has to be because of more credible historical experience.
2. Define the confidence interval width as 0.5*(upper CI ± lower CI) and either:
   - keep the confidence width fixed for each duration which leads to tighter confidence intervals as duration increases since the standard deviation of the retrospective reserve increases by duration, or
   - linearly reduce the confidence width to zero from duration 1 to the end of the coverage period.
3. Any other reasonable method could be explored.

The following is an illustration of how the prospective reserves could be adjusted for a hypothetical in force block of 20-year, fully discrete term insurance policies issued over the past 10 years. For this hypothetical illustration, we assume the following:

1. For each issue year, 100 policies are issued and they are randomly issued over issue ages 35–55 and face amounts $100,000–$500,000.
2. Policy premiums are calculated based on the actuarial equivalence principle.
3. For durations 1–5 (i.e., more recent issues), actual historical mortality is assumed to be 25% lower than reserving assumptions.
4. For durations 6–10 (i.e., earlier issues), actual historical mortality is assumed to be 25% higher than reserving assumptions.
5. Prospective reserves are adjusted based on deviations of the realized retrospective accumulated net asset random variable from the confidence interval of the retrospective accumulated net asset random variable. The confidence interval is based on 0.10*SD for policies in duration 10 at the valuation date, 0.20*SD for policies in duration 9, etc. and 1*SD for policies in duration 1 at the valuation date as illustrated in Table 5. Note that issue year 1 represents policies in duration 10, issue year 10 represents policies in duration 1, and so forth.
6. Assume the only decrement is mortality and that the prospective reserve is being calculated at end of duration 10.

Table 6 shows how the prospective reserve per $1000 is adjusted by duration to reflect actual mortality experience based on our pre-established confidence interval methodology as illustrated in Table 5.

Table 5 Retrospective accumulated net asset random variable confidence band example

Duration	Issue year	Retrospective accum net asset RV		Retrospective accum net asset RV confidence band			
		Mean	SD	Lower bound		Upper bound	
10	1	17.60	15.83	Mean − 0.1 * SD	16.02	Mean + 0.1 * SD	19.19
9	2	16.60	14.53	Mean − 0.2 * SD	13.69	Mean + 0.2 * SD	19.50
8	3	15.86	12.35	Mean − 0.3 * SD	12.16	Mean + 0.3 * SD	19.57
7	4	13.10	10.33	Mean − 0.4 * SD	8.97	Mean + 0.4 * SD	17.24
6	5	13.05	9.54	Mean − 0.5 * SD	8.28	Mean + 0.5 * SD	17.82
5	6	11.20	7.96	Mean − 0.6 * SD	6.42	Mean + 0.6 * SD	15.97
4	7	7.26	5.85	Mean − 0.7 * SD	3.16	Mean + 0.7 * SD	11.35
3	8	6.95	5.04	Mean − 0.8 * SD	2.92	Mean + 0.8 * SD	10.98
2	9	4.32	3.33	Mean − 0.9 * SD	1.32	Mean + 0.9 * SD	7.32
1	10	2.33	2.04	Mean − 1 * SD	0.29	Mean + 1 * SD	4.37

Table 6 Prospective reserve adjustment example

Duration	Issue year	Realized retro loss RV		Expected prosp loss	Adjusted prosp loss	Realized prosp loss
		Mean	Deviation	Mean (reserve)	Mean (reserve)	Mean (reserve)
10	1	13.30	−2.72	17.60	20.33	26.91
9	2	13.09	−0.60	16.60	17.20	26.04
8	3	13.00	Within interval	15.86	15.86	25.68
7	4	11.07	Within interval	13.10	13.10	22.10
6	5	11.31	Within interval	13.05	13.05	23.05
5	6	12.44	Within interval	11.20	11.20	1.12
4	7	7.94	Within interval	7.26	7.26	−0.65
3	8	7.47	Within interval	6.95	6.95	−2.54
2	9	4.55	Within interval	4.32	4.32	−4.01
1	10	2.42	Within interval	2.33	2.33	−6.13
	Aggregate	9.66		10.80	11.11	11.04

Based on these tables, we make the following observations:

- The realized retrospective accumulated net asset random variable per $1000 of insurance is simply the mean of the retrospective accumulated net asset random variable and modifying the annual mortality based on the actual historical mortality assumptions (3) and (4) above.
- The realized retrospective accumulated net asset random variable is then compared to the theoretical mean and standard deviation of the retrospective accumulated net asset random variable based on the original reserving assumptions to determine the adjustment to the prospective reserves per $1000.

We can additionally make the following observations. First, since the standard deviation of the retrospective accumulated net asset random variable varies by duration, the impact of actual mortality experience varying from reserving assumptions has to be analyzed by issue year. Second, the overall realized prospective reserve is $11.04 per $1000 of face amount. This represents the mean of the prospective loss random variable using the actual mortality assumptions of 25% lower mortality for more recent issues in durations 1–5 and 25% higher mortality for earlier issues in durations 6–10. Third, the overall realized retrospective reserve is $9.66. Based on our approach of varying confidence interval to adjusting the prospective reserves, the overall adjusted prospective reserve per $1000 of insurance is $11.11, while the overall expected prospective reserve is $10.80. This represents an overall increase in prospective reserves of 30 cents for every $1000 of insurance. Finally, as shown in Table 7, for the in force block in year 10 after annual sales of 100 policies per year, there are approximately 993 remaining policies with an aggregate face amount of $297,226,683. Then the adjusted prospective reserve results in an increase of $93,471 in aggregate prospective reserves. This translates to a $22,808 higher than the overall mean of the prospective loss random variable based on actual mortality experience (i.e., realized prospective reserve). This implies a slight degree of conservatism in our methodology for adjusting aggregate prospective reserves.

Table 7 Difference between the adjusted and expected prospective reserves

Remaining policies	993
Remaining policies face amount	297,226,683
Expected retrospective reserve	10.80
Expected prospective reserve	10.80
Adjusted prospective reserve	11.11
Realized prospective reserve	11.04
Expected aggregate prospective reserve	3,210,105
Adjusted aggregate prospective reserve	3,303,576
Realized aggregate prospective reserve	3,280,768
Per $1000 difference between expected and realized prospective reserves	−0.24
Per $1000 difference between adjusted and realized prospective reserves	0.08
Aggregate difference between expected and realized prospective reserves	(70,662)
Aggregate difference between adjusted and realized prospective reserves	22,808

4 Concluding Remarks

The implications of this paper are important for a few reasons:

1. This paper expands the actuarial literature on unlocking reserve assumptions based on the retrospective accumulated net asset random variable, a concept that is similar to the prospective loss random variable that is used to calculate reserves. Similar retrospective concept has appeared in Arias Lopez and Garrido (2001) and Ramsay (1993).
2. The retrospective accumulated net asset random variable as defined in this article has practical implications in developing a claims tracking and monitoring process for a company and in adjusting prospective reserves in a systematic manner that would satisfy Principle Based Reserving (PBR) standards. The PBR approach is being gradually adopted by the National Association of Insurance Commissioners (NAIC) for calculating more realistic reserves. See Mazyck (2013).
3. The methodology recommended in this article is timely because PBR regulation allows insurance companies to use their own experience to value life insurance reserves. The approach suggested here can also be viewed as a methodical way of tracking and monitoring insurance claims experience. See Vadiveloo et al. (2014).

The paper has focused on the retrospective accumulated net asset random variable for a term insurance product. Clearly, our findings can be extended to other insurance products like endowment insurance, whole life insurance, disability income, long term care, life annuities, and pension plan products. For disability income and long-term care, the retrospective accumulated net asset random variable provides historical information on how actual incidence and termination rates vary from expected and whether they are significant enough to adjust the prospective reserves for the business. For annuities and pension products, the retrospective accumulated net asset random variable provides insights into the longevity risk for these products and how prospective reserves may be adjusted to reflect actual longevity experience that is significantly deviating from expected.

With this paper, future students of mathematics of life contingencies may learn about the importance of a retrospective accumulated net asset random variable in assisting insurance companies provide information on historical claims experience and how prospective reserves may be adjusted to reflect this emerging actual experience. This may also help trigger their appreciation of the concept of the retrospective reserve, rather than simply mathematically demonstrating the equivalence between the retrospective and prospective reserves.

Acknowledgements The Goldenson Center wishes to acknowledge Pratish Modi, a graduate student in actuarial science at the University of Connecticut and Ruth Nieh, a junior Honors student also at the University of Connecticut who did her honors thesis on this topic, for their assistance in this research project. The authors also appreciate the comments and suggestions made by the reviewer which helped improve the final version of this paper.

References

Arias Lopez, R., Garrido, J.: Some properties and inequalities related to the *kth* inverse moment of a positive binomial variate. Revista de Matemática: Teoría y Aplicaciones **8**, 1–18 (2001)

Atkinson, D.B., Dallas, J.W.: Life Insurance Products and Finance – Charting a Clear Course. Society of Actuaries, Schaumburg, IL (2000)

Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A., Nesbitt, C.J.: Actuarial Mathematics. Society of Actuaries, Schaumburg, IL (1986)

Dickson, D.C., Hardy, M.R., Waters, H.R.: Actuarial Mathematics for Life Contingent Risks, 2nd edn. Cambridge University Press, Cambridge (2013)

Financial Accounting Standards Board: Statement of Financial Accounting Standards No. 97, Stamford, CT (1987)

Gerber, H.U.: A probabilistic model for (life) contingencies and a delta-free approach to contingency reserves. Trans. Soc. Actuar. **28**, 127–148 (1976)

Gerber, H.U.: Life Insurance Mathematics, 3rd edn. Springer, New York (1997)

Longley-Cook, A., Shaw, D., Sherrill, M., Vadiveloo, J.: Stochastic DAC unlocking for variable annuity products. The Financial Reporter, pp. 1–6 (March 2001)

Manning, Jr., C.P.: Standard valuation law. Rec. Soc. Actuar. **16**, 1105–1124 (1990)

Mazyck, R.: The future of life insurance regulation – principle-based reserves. CIPR Newsletter, pp. 11–12 (2013)

Ramsay, C.M.: A note on random survivorship group benefits. ASTIN Bull. **23**, 149–156 (1993)

Vadiveloo, J., Niu, G., Xu, J., Shen, X., Song, T.: Tracking and monitoring claims experience: a practical application of risk management. Risk Manage. **31**, 12–15 (2014)

Spatial Statistical Tools to Assess Mortality Differences in Europe

Patricia Carracedo and Ana Debón

Abstract In general, life expectancy has increased in the whole of Europe in recent decades, especially in western European countries. However, this study detected that the observed mortality is higher than expected in eastern European countries, widening the gap between these countries and Western Europe. The main objective of this paper is to study the space dependence of significant clusters through a spatial panel data model. There are many studies that address the decrease of mortality in Europe. None of them uses spatial methodology to detect significant clusters between countries with similar mortality, implementing in turn a spatial model which controls the space dependence of the European countries over time. Thus, the objective of this study is to determine differentiated behavior areas and control the spatial interaction between European countries over time applying a spatial panel data model. The methodology takes into account the neighboring relationships between the countries. The performance of the model was assessed using the methods of goodness of fit, residual variance, and determination coefficient. This statistical methodology was applied to 26 European countries over the period 1990–2009. The R free software environment for statistical computing was used to perform the whole analysis.

Keywords Mortality • Spatial cluster • Spatial panel data models • R

P. Carracedo (✉)
Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Camino de Vera, s/n 46022 Valencia, Spain

Valencian International University (VIU), C/ Gorgos, 5, 46021 Valencia, Spain
e-mail: patcarga@posgrado.upv.es

A. Debón (✉)
Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Camino de Vera, s/n 46022 Valencia, Spain
e-mail: andeau@eio.upv.es

J.A. Londoño et al. (eds.), *Actuarial Sciences and Quantitative Finance*, Springer Proceedings in Mathematics & Statistics 214, https://doi.org/10.1007/978-3-319-66536-8_3

1 Introduction

Although, in recent decades, mortality has declined in all the countries in the European Union, considerable differences in the levels of mortality between countries (Vaupel et al., 2011) are found, especially between eastern and western countries (Meslé and Vallin, 2002). European countries have suffered a situation of divergence between Eastern and Western Europe (Leon, 2011; Vaupel et al., 2011), especially after the collapse of the Soviet system. The health division between the east and west was firstly due to the clash between two areas during the twentieth century: the economic and the political; and secondly to the collapse of the Soviet Union (Vågerö, 2010). Thus, the gap in Europe starts at least in the twentieth century.

Between 1970–1984 the mortality of the communist countries of Central and Eastern Europe (CEE) such as The Czech Republic, Hungary, Poland and Slovakia and the Baltic states: Estonia, Latvia and Lithuania suffered a slow growth. The anti-alcohol campaign introduced by Michael Gorbachev over the period 1984–1987 in Russia produced an increase in life expectancy. The impact was most pronounced in the reduction of mortality due to injuries, poisoning, and some cardiovascular disease among adult males (Bobadilla et al., 1997). In 1989–1991 with the collapse of the Berlin wall, CEE countries experienced a decline in mortality (Leon, 2011) in response to politic-economic change. In contrast, Russia, as well as the poorest republics of the former Soviet Union including The Baltic States suffered an increase in mortality. At the end of 2008, the Russian Ministry of Health proposed a set of ambitious targets to improve the health of the population (Leon, 2011). The collapse of the system in 1989 trigged a health crisis. This primarily attacked communist countries, preventing their progress while western European countries began to progress due to new advances in health care, specifically in the treatment of cardiovascular diseases (Meslé and Vallin, 2002).

Therefore, mortality has not only varied with time, but has also varied depending on the country, since not all of them have the same health and economic conditions (EUROSTAT, 2009). Europe is a continent with countries which have progressed together but in a very different way, leading to the existence of great variability between their mortality rates, particularly between eastern and western countries (Meslé and Vallin, 2002). Figure 1 shows the quintiles of mortality in Europe in 1990, 2000, and 2009, quantified by means of the Standardized Mortality Ratio (SMR). This ratio will be detailed in Sect. 2.2.1. It shows that the SMR of eastern countries is higher than western countries and is growing over time.

Spatial econometrics is a subfield of econometrics dealing with spatial interaction effects among geographical units. In the last decade, the spatial econometrics literature has focused on the specification and estimation of econometric relationships based on panel data (Elhorst, 2014a). Panel data refers to data containing a number of geographical units followed over time. Panel data have more information than longitudinal and cross-sectional studies as they contain more variability, less collinearity, more degrees of freedom, and more efficiency among the variables (Baltagi, 2008). Spatial panel data models can be used to explain the behavior of geographical units if they are related to each other.

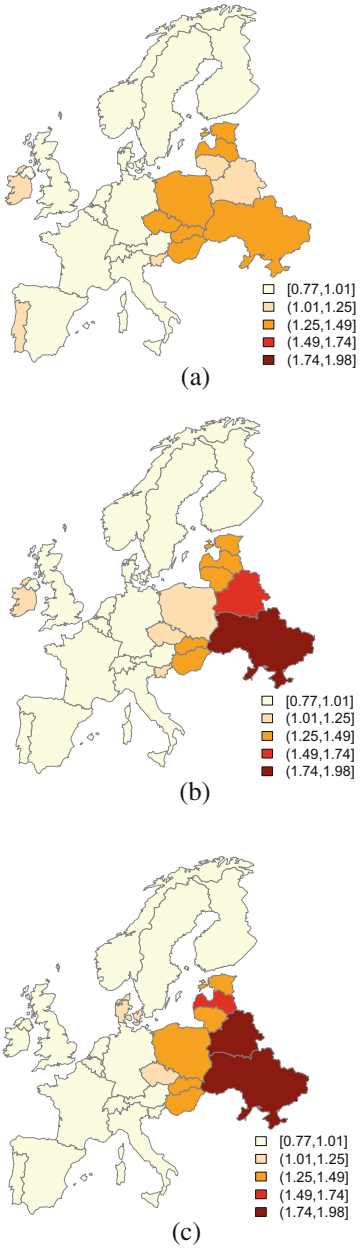


Fig. 1 Standardized mortality ratio in Europe. (**a**) 1990. (**b**) 2000. (**c**) 2009

Thus, this paper is motivated by the interest in the inequalities between the health systems in different European countries (Spinakis et al., 2011). The main objective of this paper is determine differentiated behavior zones and study the spatial interaction between the European countries over time applying a spatial method-

ology. This methodology takes into account neighboring relationships between the countries. Firstly, significant clusters of European countries with similar mortality were detected and secondly a spatial panel data model was applied to model the space dependence in the geographical units over time. The performance of the model was assessed using the well-known measure of goodness of fit named residual variance (σ^2) and determination coefficient (R^2).

This paper is structured as follows: Sect. 2 starts by describing the database of the selected countries. The section continues by detailing the spatial methodology which was used to identify clusters of countries with similar mortality and ends with an exposition of the spatial panel data implemented. In Sect. 3, the main results of applying the spatial method from the previous section to the database are shown. And finally, Sect. 4 presents the main conclusions obtained in this study.

2 Material and Methods

2.1 Data

This study deals with mortality data for European countries for the period between 1990 and 2009 and for an age range from 0 to 110+ considering a country as the unit of analysis. Data were taken from the Human Mortality Database (2014) for a total of 26 countries: Austria, Belarus, Belgium, The Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, The Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland, The United Kingdom, and Ukraine. These 26 countries were considered as they have common information in the database for the maximum time range 1990–2009 and for an age range from 0 to 110+.

With the aim of explaining the behavior of mortality depending on demographic and economic variables (Cutler et al., 2006), information about five variables for these 26 countries and 20 years were collected from The World Bank Database (2015). These variables were: population growth, gross domestic product (GDP), birth rate, activity rate, and road sector energy consumption.

- *Population growth (annual %)*: The annual population growth rate for year t is the exponential rate of growth of the midyear population from year $t-1$ to t, expressed as a percentage. Population is based on the fact definition of population, which counts all residents regardless of legal status or citizenship.
- *Gross domestic product (annual %)*: Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2010 U.S. dollars. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural

resources. This variable contains missing data. The function *knnImputation* of the library *DMwR* (Torgo, 2010) was used to impute the missing data. This function fills the missing values with a local weighted average.

- *Crude Birth rate (per 1000 people)*: Crude birth rate indicates the number of live births occurring during the year per 1000 population estimated at midyear. Subtracting the crude death rate from the crude birth rate provides the rate of natural increase, which is equal to the rate of population change in the absence of migration.
- *Activity rate (% of total population ages 15+)*: Labor force participation rate is the proportion of the population aged 15 and older that is economically active: all people who supply labor for the production of goods and services during a specified period.
- *Road sector energy consumption (% of total energy consumption)*: Road sector energy consumption is the proportion of total energy used in the road sector including petroleum products, natural gas, electricity, and combustible renewable and waste.

In order to avoid collinearity in the covariates, the Variance Inflation Factor (VIF) was taken into account. The covariates with values of VIF less than 2 were selected. Lastly, these variables were: GDP, activity rate, road sector energy consumption, and birth rate.

Statistical analysis was performed using the software R Core Team (2015) together with some R-packages: demography (Hyndman et al., 2014), maptools (Bivand and Lewin-Koh, 2014; Charpentier, 2014), spdep (Bivand, 2012; Charpentier, 2014), GeoXp (Laurent et al., 2012), rgdal (Bivand et al., 2016), Gmisc (Gordon, 2016), RColorBrewer (Neuwirth, 2014), splm (Millo and Piras, 2012), plm (Croissant et al., 2008), and DMwR (Torgo, 2010).

2.2 *Clustering Spatial Methodology*

Several statistics will be described in order to quantify the spatial relationship of mortality and detect groups of European countries with similar mortality.

2.2.1 Standardized Mortality Ratio (SMR)

Quantifying mortality is important to outline the epidemiological, demographic, and development levels of a country. Mortality is influenced by factors such as diseases of a random nature or natural disasters, so mortality is variable over time, that is, it does not remain uniform. Public health professionals are constantly faced with comparing mortality between different geographical areas. There is no problem comparing mortality rates if populations are distributed similarly with respect to

other factors such as age, race, social class, etc. but this does not happen. When comparing mortality rates between different geographical areas, these rates will be influenced by the proportion of subjects in each age group in each geographic area. To solve this problem, standardization methods were developed. Standardization allows comparisons without the effects of differences in the size of the sub-groups of the population. There are two methods of standardization: direct and indirect (Fleiss et al., 2013). In this paper, to quantify mortality, the Standardized Mortality Ratio (SMR) was used as it is the most widely used index to compare mortality between different areas (Hinde, 1998). The indirect method produces the SMR. This ratio was obtained for each of the 26 European countries during the period 1990–2009.

Standardized Mortality Ratio (SMR) is a well-known index which compares observed deaths and expected deaths, both measured at the same moment in time. The SMR is defined as the number of deaths that would be expected in a studied population if the age specific mortality rates were those of the standard population. Its calculation is expressed as

$$SMR_{i,t} = \frac{O_{i,t}}{E_{i,t}} \quad \text{for i} \in \{1, \ldots, N\} \quad \text{and t} \in \{1, \ldots, T\} \tag{1}$$

where i is the country and t is the year. $O_{i,t}$ represents the number of observed deaths for each country i in the year t, and $E_{i,t}$ corresponds to the number of deaths in each country i in the set of European countries in the year t under the hypothesis that all the countries have the same mortality as the set of European countries. A ratio greater than 1 indicates that more mortality was observed than would have been expected, in this case there are "excess deaths." On the contrary, there are "deficit deaths" if the SMR is less than 1, a situation that occurs when there is a lower number of observed deaths than expected (Hinde, 1998).

If x is the age of death, $O_{i,t}$ is obtained as

$$O_{i,t} = \sum_{x=0}^{110+} m_{x,i,t} * P_{x,i,t} \quad \text{for x} \in \{0, \ldots, 110+\}$$

where $m_{x,i,t}$ represents the death rate and $P_{x,i,t}$ the size of the studied population at age x, country i and year t. The expected deaths can be obtained as

$$E_{i,t} = \sum_{x=0}^{110+} E_{x,i,t} \quad \text{for x} \in \{0, \ldots, 110+\},$$

where $E_{x,i,t}$ is,

$$E_{x,i,t} = m_{x,t} * P_{x,i,t},$$

and $m_{x,t}$ is the death rate at age x for the year t in the set of European countries.

$$m_{x,t} = \frac{\sum_{i=1}^{N} O_{x,i,t}}{\sum_{i=1}^{N} P_{x,i,t}} \quad \text{for i} \in \{1, \ldots, N\}.$$

2.2.2 Global Moran Index

The Global Moran's I is a summary measure that shows the intensity of the spatial dependence of all the countries in the study (Moran, 1950a,b). Positive index values indicate positive spatial autocorrelation in the European countries; when the SMR of a countries increase or decrease, the SMR of its neighbors also increases or decreases, respectively. In contrast, negative values of this index indicate negative spatial autocorrelation in European countries; when the SMR of countries increases or decreases, the SMR of its neighbors decreases or increases, respectively. Values of the index close to zero indicate the absence of spatial autocorrelation between these 26 European countries. The expression of the index is as follows:

$$GM_t = \frac{N \sum_i \sum_j W_{i,j}(SMR_{i,t} - \overline{SMR_t})(SMR_{j,t} - \overline{SMR_t})}{\sum_i \sum_j W_{i,j} \sum_i (SMR_{i,t} - \overline{SMR_t})^2} \quad \text{for i} \in \{1, \ldots, N\},$$

$$\text{j} \in \{1, \ldots, N\} \quad \text{and i} \neq \text{j}$$

where N is the total number of European countries, $\overline{SMR_t}$ is the average of the SMR in all the countries at time t, a $W_{i,j}$ is the spatial weights matrix where i and j are two different countries in the set of N European countries considered. In this study, two countries are considered neighbors when they share a border (first order of neighborhood). Only, the first order in the neighborhood structure was used (Anselin, 1995), given the importance of borders in Diehl (1992) which is a study of international conflict. The first order of neighborhood takes into account only the influence of neighbors and not the influence of the neighbors of the neighbors (second order) or the influence of the neighbors of the neighbors of the neighbors (the third order), and so on. With this, $W_{i,j}$ can take the following values:

$$W_{i,j} = 0, \quad \text{if } j \notin V(i);$$

$$W_{i,j} = \frac{1}{n_i}, \quad \text{if } j \in V(i), \text{ with } n_i = \#V(i);$$

$$W_{i,i} = 0, \quad i = 1, .., N.$$

where n_i is the number of neighbors i and $V(i)$ is the set of neighbors of a country i. When $W_{i,j} = 0$ the countries i and j are not considered neighbors, while if $W_{i,j} \neq 0$

the countries i and j are considered neighbors with a weight $1/n_i$. The total of each row is 1 because the weights $W_{i,j}$ are standardized. In the spatial weights matrix a country cannot be its own neighbor then, $W_{i,i} = 0$.

Global Moran's test for spatial autocorrelation was calculated using the R-package spdep by Bivand (2012). The result of the contrast provides the following output: the value of the observed Moran's I, its expectation which means the expected value of Moran's I under the null hypothesis of no spatial autocorrelation, its variance and the *p*-value of the Moran's test. These results may be checked against those of the Monte Carlo test. The Monte Carlo test uses random permutations of $SMR_{i,t}$ for the spatial weights matrix, to establish the rank of the observed statistic in relation to the 999 simulated values. The result of the contrast provides the following output: the value of the observed Moran's I, the rank of the observed Moran's I, and the *p*-value of the Monte-Carlo test. Both tests are sensitive to the form of the spatial weights matrix.

2.2.3 Local Moran Index

The Local Moran's I is a Local Indicator of Spatial Association (LISA), which was introduced by Anselin (1995). This ratio determines whether the spatial correlation scheme detected in all countries of the study is also maintained locally. In the notation which has become usual in this context, L denotes SMR values of a country that are Lower (L) than its mean and H denotes SMR values of a country that are Higher (H) than its mean. In the same way for the neighbors, L and H denote mean SMR of neighbors that are Lower (L) or Higher (H) than its mean, respectively. Thus, each observation could be placed in one of four categories, as summarized in Table 1.

When the index is significant, two types of clusters are detected:

- A positive Local Moran's I indicates *Spatial Clusters* of countries with high values of SMR surrounded by neighbors also with high values of SMR, denoted by HH or spatial clusters of countries with low SMR values surrounded by neighbors also with low SMR values denoted by LL.
- A negative Local Moran's I indicates *Outlier Clusters* of countries with low SMR values surrounded by neighbors with high SMR values, denoted by LH or outlier clusters of countries with high SMR values surrounded by neighbors with low SMR values denoted by HL.

Table 1 Lisa Classifications for each country and its neighborhood

Class	Country'SMR	Neighbors'$\overline{SMR}$
HH	Above average	Above average
HL	Above average	Below average
LH	Below average	Above average
LL	Below average	Below average

The expression of the index is as follows:

$$LM_{i,t} = \frac{(SMR_{i,t} - \overline{SMR_t})}{S^2(SMR_t)} \sum_i \sum_j W_{i,j}(SMR_{j,t} - \overline{SMR_t}) \quad \text{for i} \in \{1, \ldots, N\},$$

$$\text{j} \in \{1, \ldots, N\} \quad \text{and i} \neq \text{j}$$

where $S^2(SMR_t)$ is the variance of SMR_t at time t.

Spatial clusters, sometimes referred to as hot spots, may be identified as those locations or sets of contiguous locations for which the Local Moran Index is significant. Local Moran Index can be used as the basis for a test on the null hypothesis of no local spatial association. This test gives an indication of the extent of significant local spatial cluster of similar values around one country *i* (Anselin, 1995). The sum of all Local Moran Index for all countries is proportional to the Global Moran Index. To carry out an adjustment to Local Moran Index' *p*-values based on the number of neighbors of each region the Bonferroni correction was used in this paper. The Bonferroni method is a stricter criterion for the significance level of Local Moran's I. This method stresses the *p*-values obtained for the local Moran's I. These adjusted *p*-values are dependent on the number of neighbors of a country *i* (Anselin, 1995).

2.3 Spatial Panel Data Models

Once you itemize techniques to study the spatial dependence of mortality data from 26 European countries during the period 1990–2009 in Sect. 2.2, the next step is to implement a spatial panel data model to model the space dependence of these countries during the considered period of time.

2.3.1 Panel Data

Panel data are spatial observations (regions, countries, families, households, etc.) followed with time. Therefore, panel data are a combination of two dimensions: space and time (Wooldridge, 2010).

Normally, panel data are distinguished from others by the spatial and temporal extension of the data, the various types of panels are

- *Micro Panels*: Those panels with more cross observations than periods.
- *Macro Panels*: Those panels with less cross observations than periods.
- *Random Field Panels*: They are panels with a very wide temporal and transverse dimension.

Depending on the existence or absence of missing data, panel data can be of two types

- *Balanced or Complete Panels*: If all the units studied are observed throughout the study period.
- *Imbalanced or Incomplete Panels*: If there are missing data, the time period varies between individuals.

Specifically, data from this study are a *Micro Panel*, since there are more space observations (26 countries) than periods of time (20 years) and a *Balanced Panel* because there aren't missing data in any variable (the values of GDP were imputed).

2.3.2 Spatial Models for Panel Data: Spatial Lag Model with Fixed Effects

Currently, spatial econometrics is emphasizing the specification and estimation of econometric relationships based on information panels containing data. This interest can be explained by the increased availability of a large amount of data in which the spatial units (municipalities, regions, states, countries, postal codes, etc.) are followed over time. Panel data also allow for the specification of more complicated behavioral hypotheses, including effects that cannot be addressed using pure cross-sectional data. Panel data offer researchers extended modeling possibilities as compared to the single equation cross-sectional setting, which was the primary focus of the spatial econometrics literature for a long time (Elhorst, 2014a).

A panel data model is a regression model which uses the temporal and spatial heterogeneity of the panel structure to estimate parameters of interest (Elhorst, 2014b). This model differs from cross-section regression or time series in that it considers both the spatial and temporal dimension, which favors study, especially in periods of great change. Panel data models offer advantages over cross-section regression or time series. They control unobserved heterogeneity produced by both spatial and temporal units which reduces the problems of multicollinearity between the variables (Kennedy, 2003). The geographical units are observed over time and this fact cannot be studied using purely cross-sectional or time series studies. Panel data usually contain more degrees of freedom and more sample variability than cross-sectional data as time series data, hence improving the efficiency of econometric estimates (Hsiao et al., 2002). Some problems appear using panel data. First of all design and data collection problems are more complicated than in the case of cross-sectional data or time series (Arbia and Piras, 2005).

Elhorst (2014a) provides a review of the spatial panel data models most commonly used in research

I Spatial Lag Model with Fixed Effects (SLMFE).
II Spatial Error Model with Fixed Effects (SEMFE).
III Spatial Lag Model with Random Effects (SLMRE).
IV Spatial Error Model with Random Effects (SEMRE).

The typology of our data leads us to implement a [I] Spatial Lag Model with Fixed Effects. The reasons are:

- Fixed Effects: The fixed effects model is generally more appropriate than the random effects model since spatial econometricians tend to work with space-time data of adjacent spatial units located in unbroken study areas (Elhorst, 2014a). In addition, it attempts to model the behavior of each country and time individually. This model, known as two-ways, assumes that differences between countries and time are constants (Asteriou and Hall, 2015). For this reason, spatial and temporal dummy variables are incorporated, which model the unobserved characteristics of cross-sectional units (not changing over time but affecting the dependent variable, examples of these characteristics are religion, sex, education, etc.) and the unobserved characteristics of temporal units (not changing with countries but affecting the dependent variable, for example a great depression, world war, etc.).
- Spatial Lag: The value of the SMR in a country depends on the value of the SMR in another adjacent country. This fact will be confirmed in the next Sect. 3.1.

Then the SLMFE is defined mathematically as

$$y_{it} = \alpha + \lambda \sum_{j=1}^{N} W_{i,j} y_{jt} + X_{it}\beta + \mu_i + \nu_t + \varepsilon_{it} \tag{2}$$

where:
i represents the countries;
t represents the years;
y_{it} represents a vector of dimension $NT \times 1$ corresponding to observations of the dependent variable for each country i and year t;
α is the mean intercept. This will be estimated with the condition that the sum of the spatial and temporal effects is zero (Hsiao, 2014). In this way the spatial effect represents the deviation of the spatial unit i from the mean α and the time effect represents the deviation of the time unit t from the mean α;
λ is the spatial parameter associated with the dependent variable;
$W_{i,j}$ spatial weights matrix where i and j represent whichever two of the N countries of dimension $N \times N$;
X_{it} is a matrix of dimension $NT \times K$ of observations on the independent variables;
β vector of dimension $K \times 1$ of fixed but unknown parameters corresponding to observations of the independent variables;
μ_i is the spatial fixed effect (not spatially autocorrelated) which captures the unobservable characteristics that change across countries but remain constant over time;
ν_t is the temporal fixed effect (not temporally autocorrelated) which captures the unobservable characteristics that change over time but remain constant across countries;

ε_{it} is a vector of independent and identically distributed error terms (not spatially autocorrelated) of dimension $NT \times 1$ which captures the unobservable characteristics that change over time and across countries.

3 Results

In this section the results of applying the methodology exposed in Sect. 2.2 and Sect. 2.3 are considered. For the sake of brevity, only the maps corresponding to the years 1990, 2000, and 2009 are shown here. Readers interested in maps for all the years can request them from the authors.

3.1 Clustering Spatial Methodology

The results of measuring mortality in Europe in terms of the SMR, calculated according to the expression (1) are shown in this section. To quickly compare the mean and variance of the SMR of each European country studied between 1990 and 2009 Fig. 2 was prepared. It shows the variability of the SMR of each country in the studied period. For example, Belarus and Ukraine (eastern countries) are the countries with more variability in SMR, whereas Belgium and Spain (western countries) are two of the countries with minor variability in the SMR.

In addition, the countries with mean SMR values higher and lower than 1 are identified:

- Countries with values of SMR higher than 1: Belarus, Slovakia, Estonia, Hungary, Latvia, Lithuania, Poland, The Czech Republic, and Ukraine. In this case these countries suffer "excess deaths" in the studied population, because the observed deaths in these countries are higher than the deaths that would be expected if they behaved similarly to the set of European countries. These levels of SMR above 1 were maintained during the period of study, and therefore, these countries are not considered privileged.
- Countries with values of SMR lower than 1: Germany, Austria, Belgium, Spain, Finland, France, The Netherlands, Italy, Luxembourg, Norway, Portugal, The United Kingdom, Sweden, and Switzerland. In this case these countries suffer "deficit deaths" in the studied population, because the observed deaths in these countries are lower than the deaths that would be expected if they behaved like the set of European countries. These levels of SMR lower than 1 were maintained during the period of study, and therefore, these countries are considered privileged.

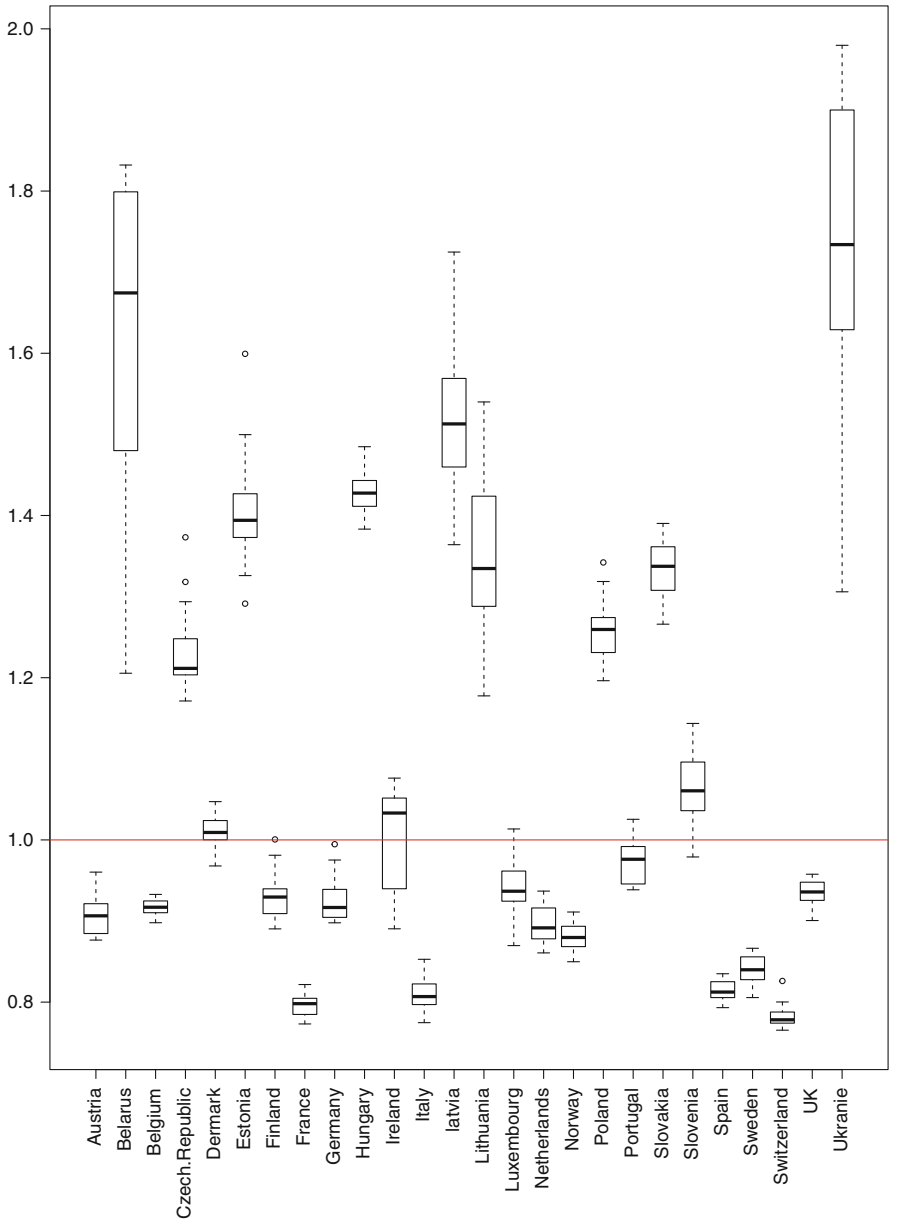


Fig. 2 Box plot of the SMR for each country

- Countries with SMR values of around 1: Denmark, Ireland, and Slovenia. These countries have similar deaths to what would be expected if they behaved as the set of European countries between the years studied, and they have a status similar to the general one.

It confirms that the observed mortality is higher than expected in the eastern countries over time.

Figure 3 shows Moran scatter plots where the SMR value for a country is plotted against the average SMR of its neighbors for years 1990, 2000, and 2009 respectively. All the graphs obtained for all years indicate that there is a positive spatial correlation in the set of European countries. The countries that move away from the central trend are marked: Portugal (PT), Estonia (EE), Belarus (BY), Lithuania (LT), and Ukraine (UA).

In order to confirm the presence of spatial autocorrelation, the null hypothesis $H_0 : GM_t = 0$ is tested. The p-value is obtained using asymptotic distribution or by means of a Monte Carlo test (Bivand, 2012). The results of Moran and Monte Carlo tests for the considered period are shown in Table 2 which includes the value of the observed Moran's I, its expectation and variance, the p-value of the Moran's test (M) and in the last column the p-value of the Monte Carlo test (MC). The p-values obtained for all years are significant (p-values <0.05), indicating that there is a spatial dependence in the observed mortality.

As stated in Sect. 2.2.3, significant values of the Local Moran Index show two types of clusters: *Spatial clusters* and *Outlier clusters*. Maps in Fig. 4 show spatial clusters of type HH (high SMR) and LL (low SMR) which means that the observed mortality in the countries belonging to different clusters is similar enough to be able to form these clusters. In addition, the cluster center is identified. The center cluster is unique and represents the country located in the middle of the cluster. When several center clusters appear inside a single cluster it means that there are several clusters belonging to bordering countries which form a single macrocluster.

In Fig. 4 two significant clusters of different European countries are observed until the year 2002, identifying the center and neighbors in these: a cluster of high SMR consisting of Eastern European countries (Lithuania, Latvia, Estonia, Ukraine, Belarus, Slovakia, Hungary, and Poland) and another cluster of low SMR consisting of Western European countries (Spain, Italy, France, Switzerland, Germany, Luxembourg, and Belgium), as the Local Moran's I significant values indicate. Non-significant values of the Local Moran Index therefore identify countries which do not belong to any cluster (The Czech Republic, United Kingdom, Denmark, Finland, Ireland, The Netherlands, Norway, Slovenia, Portugal, Sweden and Austria).

It is important to emphasize that the center of cluster LL (low SMR) is France. This is unique and remains constant over the period 1990–2009. This cluster LL disappears from 2002 because the variability of the mortality in western countries has been increasing since 2002. On the contrary there are several clusters of type HH (high SMR) which form a unique macrocluster HH. For this reason in the cluster HH (high SMR) several center clusters are observed. These center clusters differ over the same period, moving from west to east of Europe.

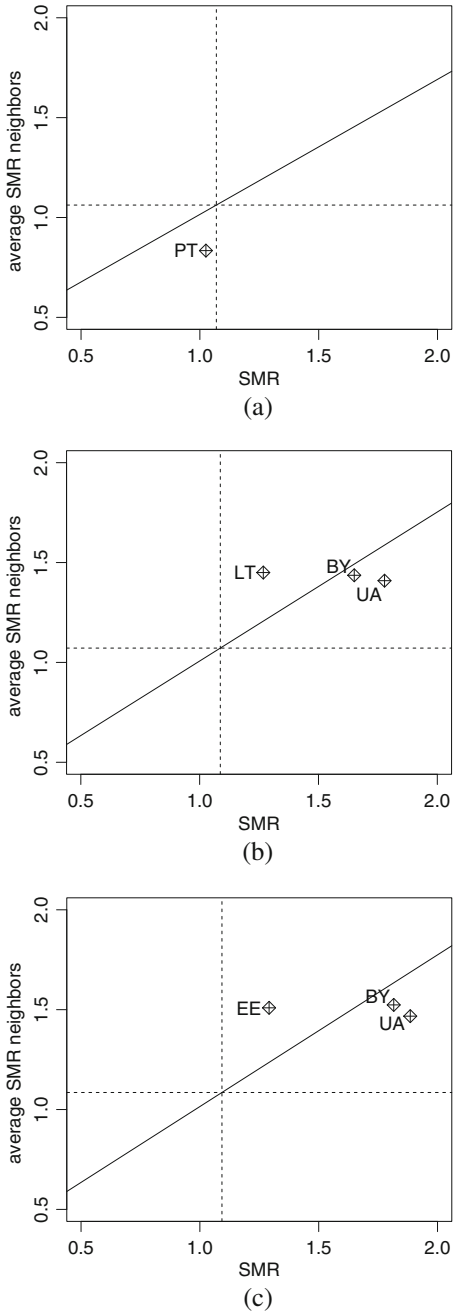


Fig. 3 Behavior of the Global Moran's Index in Europe. (**a**) 1990. (**b**) 2000. (**c**) 2009

Table 2 Values of the Global Moran's I and associated *p*-values to SMR

Year	I.Moran	Expectation	Variance	*p*-value M	*p*-value MC
1990	0.680	−0.040	0.03	0.000	0.001
1991	0.738	−0.040	0.03	0.000	0.001
1992	0.748	−0.040	0.03	0.000	0.001
1993	0.792	−0.040	0.03	0.000	0.001
1994	0.818	−0.040	0.03	0.000	0.001
1995	0.798	−0.040	0.03	0.000	0.001
1996	0.772	−0.040	0.03	0.000	0.001
1997	0.774	−0.040	0.03	0.000	0.001
1998	0.790	−0.040	0.03	0.000	0.001
1999	0.759	−0.040	0.03	0.000	0.001
2000	0.746	−0.040	0.03	0.000	0.001
2001	0.761	−0.040	0.03	0.000	0.001
2002	0.737	−0.040	0.03	0.000	0.001
2003	0.733	−0.040	0.03	0.000	0.001
2004	0.749	−0.040	0.03	0.000	0.001
2005	0.748	−0.040	0.03	0.000	0.001
2006	0.769	−0.040	0.03	0.000	0.001
2007	0.769	−0.040	0.03	0.000	0.001
2008	0.748	−0.040	0.03	0.000	0.001
2009	0.759	−0.040	0.03	0.000	0.001

It is striking that Austria belongs to the cluster HH (high SMR) consisting of eastern countries in the years 1990 and 1991. This is because Austria in these years had a common border with Slovakia and Hungary, countries belonging to the center cluster.

3.2 Spatial Lag Model with Fixed Effects (SLMFE)

The splm R-package by Millo and Piras (2012) is used to estimate the SLMFE model. In the splm function the formula of the model, the data, spatial weights matrix and the type of model have to be specified. The dependent variable in the model is the logarithm of SMR, because the SMR has a strong asymmetry. The independent considered variables are: GDP, activity rate, road sector energy consumption, and birth rate.

The results of the fitted SLMFE with the four covariates are shown in Table 3, which include the parameters of the model, the estimated value of these parameters, the standard error, and the *p*-values or significance associated with each of the parameters.

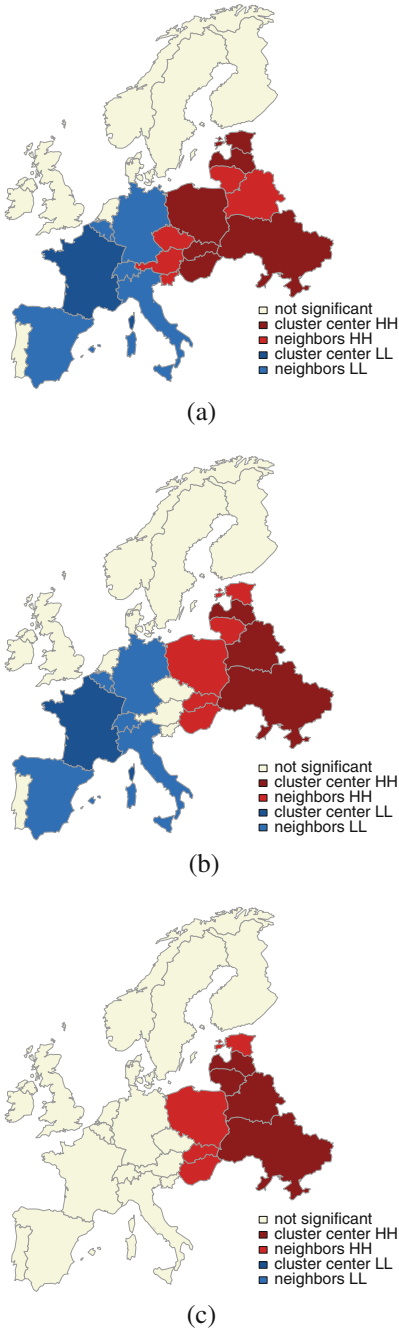


Fig. 4 Clusters map in Europe. (**a**) 1990. (**b**) 2000. (**c**) 2009

Table 3 Output of SLMFE model with four covariates

Parameters	Estimate	Standard error	t-value	p-value
α	0.2810	0.0525	5.3527	0.0000***
λ	0.3522	0.0392	8.9933	0.0000***
β_{GDP}	0.0042	0.0006	6.9958	0.0000***
$\beta_{activity\ rate}$	−0.0037	0.0009	−3.9394	0.0000***
$\beta_{road\ energy\ cons}$	−0.0042	0.0008	−5.4381	0.0000***
$\beta_{birth\ rate}$	0.0028	0.0024	1.1919	0.2333

***p-values <0.05 are significant

Table 4 Output of SLMFE model with three covariates

Parameters	Estimate	Standard error	t-value	p-value
α	0.2896	0.0523	5.5337	0.0000***
λ	0.3465	0.0391	8.8624	0.0000***
β_{GDP}	0.0041	0.0006	6.9201	0.0000***
$\beta_{activity\ rate}$	−0.0033	0.0009	−3.7814	0.0002***
$\beta_{road\ energy\ cons}$	−0.0041	0.0008	−5.3165	0.0000***

***p-values < 0.05 are significant

Only the variable birth rate is non-significant with a p-value>0.05, therefore it was removed from the model (Table 3).

Table 4 shows the result of the fitted SLMFE with three significant covariates. In this model all covariates are significant and had the expected sign. The positive sign of variable GDP is noteworthy. This result is consistent with EUROSTAT (2013) which shows that although health conditions are related to GDP, they are not completely dependent on the production of wealth in a given economy. The differences between countries can also be attributed to other factors as the quality of healthcare services, if the organizations are private or public, environmental factors, and cultural choices. These factors affect health outcomes (EUROSTAT, 2013).

The parameter α represents the average value of the fixed effects in the SLMFE model. All covariates are significant which means that the variables GDP, activity rate and road sector energy consumption are important to explain the logarithm of SMR. Variations in these three covariates cause variations in the logarithm of the SMR of a country and in turn in the value of the logarithm of the SMR in the connected countries. Moreover, the estimate for the spatial parameter (λ) is positive (0.3465) and highly statistically significant (p-value=0 <0.05). This indicates that, besides the contributions that the covariables realize to the logarithm of the SMR of a country its value increases by 34.65% when on average the logarithm of the SMR values corresponding to the environment also increases.

The results of the estimation of spatial fixed effects of SLMFE are shown in Table 5, which include the spatial fixed parameters of the model, the estimated value of these parameters μ_i, the standard error and the p-values or significance associated with each of the parameters. The value of μ_i represents the deviation of

Table 5 Estimation of spatial effects SLMFE model with three covariates

Country	Estimate μ_i	Standard error	t-value	p-value
Austria	−0.1368	0.0534	−2.5632	0.0104*
Belgium	−0.1182	0.0462	−2.5564	0.0106*
Belarus	0.2724	0.0529	5.1495	0.0000***
Switzerland	−0.1821	0.0608	−2.9974	0.0027**
Czech Republic	0.1167	0.0532	2.1946	0.0282*
Germany	−0.1025	0.0523	−1.9614	0.0498*
Denmark	0.0349	0.0592	0.5886	0.5561
Estonia	0.1350	0.0551	2.4527	0.0142*
Spain	−0.1978	0.0496	−3.9840	0.0000***
Finland	−0.0673	0.0545	−1.2360	0.2164
France	−0.2256	0.0500	−4.5154	0.0000***
Hungary	0.2044	0.0451	4.5344	0.0000***
Ireland	−0.0074	0.0541	−0.1371	0.8909
Italy	−0.2123	0.0452	−4.6915	0.0000***
Lithuania	0.1161	0.0544	2.1342	0.0328*
Luxemburg	0.0224	0.0564	0.3972	0.6912
Latvia	0.2544	0.0545	4.6665	0.0000***
Netherlands	−0.1217	0.0550	−2.2123	0.0269*
Norway	−0.1231	0.0575	−2.1433	0.0321*
Poland	0.0522	0.0511	1.0211	0.3072
Portugal	0.0381	0.0554	0.6889	0.4909
Sweden	−0.1727	0.0563	−3.0659	0.0022**
Slovenia	0.0213	0.0528	0.4042	0.6860
Slovakia	0.1271	0.0541	2.3502	0.0188*
Ukraine	0.3612	0.0513	7.0437	0.0000***
Uk	−0.0887	0.0557	−1.5932	0.1111

*p-values < 0.05, **p-values < 0.01, ***p-values < 0.001 are significant

country i from the intercept α. The estimations of spatial effects with a negative sign belong to the western countries inside the cluster of low SMR. This means that in these countries the unobserved characteristics affect the logarithm of the SMR in a negative form, compared with the average of all countries. On the contrary, the estimations of spatial effects with a positive sign belong to the eastern countries inside the cluster of high SMR. This means that in these countries the unobserved characteristics affect the logarithm of the SMR in a positive form, compared with the average of all countries. Most countries with non-significant spatial effects do not form a cluster in Fig. 4.

The results of the estimation of temporal fixed effects of SLMFE are shown in Table 6, which include the temporal fixed parameters of the model, the estimated value of these parameters ν_t, the standard error and the p-values or significance associated with each of the parameters. Then the value of ν_t represents the deviation of year t from the intercept α.

Table 6 Estimation of time effects SLMFE model with three covariates

Year	Estimate ν_t	Standard error	t-value	p-value
1990	−0.0074	0.0532	−0.1392	0.8893
1991	−0.0042	0.0528	−0.0800	0.9363
1992	0.0030	0.0525	0.0565	0.9549
1993	0.0081	0.0519	0.1551	0.8767
1994	−0.0010	0.0522	−0.0190	0.9849
1995	−0.0081	0.0522	−0.1541	0.8775
1996	−0.0134	0.0522	−0.2567	0.7974
1997	−0.0153	0.0528	−0.2900	0.7718
1998	−0.0060	0.0527	−0.1135	0.9096
1999	−0.0038	0.0526	−0.0720	0.9426
2000	−0.0089	0.0532	−0.1673	0.8672
2001	0.0003	0.0526	0.0054	0.9957
2002	0.0033	0.0527	0.0618	0.9507
2003	−0.0025	0.0529	−0.0464	0.9630
2004	−0.0032	0.0534	−0.0591	0.9529
2005	−0.0036	0.0535	−0.0676	0.9461
2006	0.0010	0.0541	0.0178	0.9858
2007	0.0042	0.0544	0.0770	0.9386
2008	0.0130	0.0535	0.2427	0.8083
2009	0.0446	0.0518	0.8615	0.3889

Table 7 Output of Lagrange multiplier test for both effects, spatial effects, and time effects

Lagrange multiplier test	Chi square	Degrees freedom	p-value
Spatial effects	4168.5	1	0.0000***
Time effects	2.3718	1	0.1235
Spatial and Time effects	4170.9	2	0.0000***

***p-values <0.001 are significant

All the ν_t are not significant. Before excluding non-significant effects from the model, the Lagrange Multiplier test by Breusch and Pagan (1980) in the R-package plm (Croissant et al., 2008) was used. The aim of this test is to contrast the incorporation of spatial effects or time effects or spatial and time effects in the model. If the p-value is less than 0.05 the null hypothesis will be rejected and therefore it is necessary to include the considered fixed effects in the model. The results of the Lagrange Multiplier tests are showed in Table 7 which include the value of the statistic used in the contrast, degrees freedom and the p-values or significance associated with each of the tests. The Lagrange Multiplier tests conclude that the spatial and time effects cannot be excluded from the model because its p-value is the most significant.

In addition, these ν_t were represented graphically (Fig. 5) and it can be seen that time effects follow a growing trend except for the period 1994–1998. This

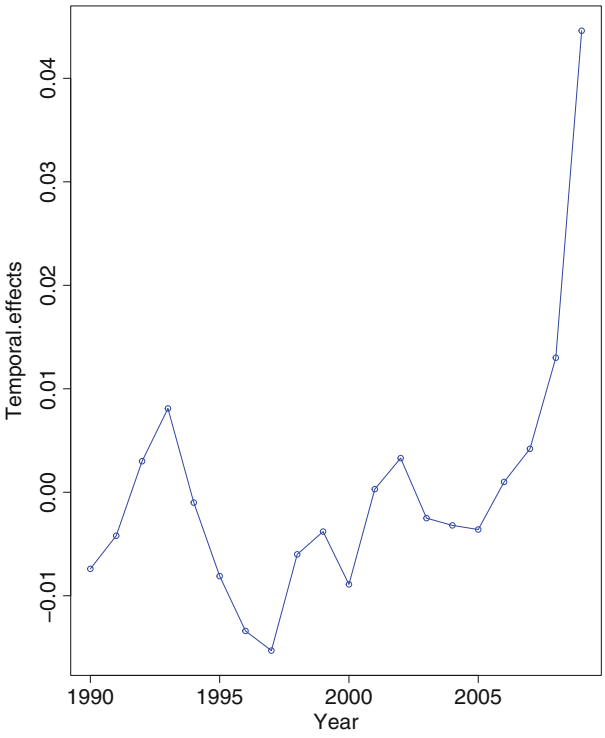


Fig. 5 Graphical representation of time effects

unfavorable evolution picks up the collapse of the Soviet system. Russia had the worst life expectancy in the year 1994, as well as Estonia, Latvia, and Lithuania. From 1994 until 1998, life expectancy in the Baltic Republics and Russia became more favorable. This was a sign of adjustment to the new circumstances (Vågerö, 2010).

Finally, the measures of goodness of fit used to validate the model were the residual variance (σ^2) and determination coefficient (R^2). Both measures indicate that the SLMFE is a good model because the value of σ^2 is very low (0.001) compared with the total variance of the model (0.057) and the value of R^2 is very high (0.97).

To check if the SLMFE model explains the spatial dependence of the 26 European countries detected by the Global and Local Moran Index, the residuals of the SLMFE model were studied. Table 8 shows the result of applying the Moran and Monte Carlo tests to the residuals of SLMFE model in the considered period. The p-values obtained for all years are not significant (p-values >0.05), indicating that the SLMFE model controls the spatial dependence.

The residuals of the SLMFE model in Fig. 6 indicate homoscedasticity because their behavior is around 0.

Table 8 Values of the Global Moran's I and associated *p*-values to residuals of SLMFE model

Year	I.Moran	Expectation	Variance	*p*-value M	*p*-value MC
1990	0.131	−0.040	0.025	0.137	0.123
1991	−0.077	−0.040	0.025	0.592	0.582
1992	0.053	−0.040	0.018	0.248	0.243
1993	0.115	−0.040	0.027	0.174	0.200
1994	−0.134	−0.040	0.017	0.767	0.791
1995	−0.057	−0.040	0.018	0.550	0.570
1996	−0.273	−0.040	0.027	0.921	0.928
1997	0.090	−0.040	0.026	0.208	0.198
1998	−0.167	−0.040	0.018	0.824	0.839
1999	0.045	−0.040	0.026	0.300	0.287
2000	−0.041	−0.040	0.023	0.503	0.498
2001	0.073	−0.040	0.027	0.245	0.235
2002	0.186	−0.040	0.027	0.084	0.088
2003	−0.095	−0.040	0.024	0.637	0.627
2004	−0.021	−0.040	0.025	0.453	0.455
2005	0.342	−0.040	0.025	0.008	0.008
2006	0.250	−0.040	0.024	0.031	0.047
2007	−0.111	−0.040	0.028	0.665	0.650
2008	0.045	−0.040	0.027	0.302	0.308
2009	−0.027	−0.040	0.022	0.465	0.466

4 Conclusions

Over a four-decade period, a health gap has opened up between European countries, in particular between the east/west (Vågerö, 2010). The gap is growing larger and has not attracted as much attention as it deserves (Leon, 2011); therefore, a deep study of the mentioned differences is necessary.

This paper quantifies and compares the mortality in Europe using the SMR. To study spatial dependence in the 26 European countries during the period 1990–2009 the Moran Global Index was used and to detect significant clusters of countries with similar mortality the Local Moran Index was used. These last two measures are used in fields such as epidemiology, demography, and econometrics. In addition, once spatial dependence was confirmed, a spatial panel data model was implemented to control the space dependence in the European countries over time.

From the results as regards countries clustering described in Sect. 3.1, the main conclusions regarding the quantification of mortality and detection of spatial clusters between European countries are the following.

The SMR remains higher than 1 over time in eastern European countries, while in the rest of Europe, the SMR is less than 1. These results are consistent with those obtained in papers such as Vaupel et al. (2011) and Mackenbach et al. (2013), where the countries of Eastern Europe have a very low life expectancy compared to the rest of Europe.

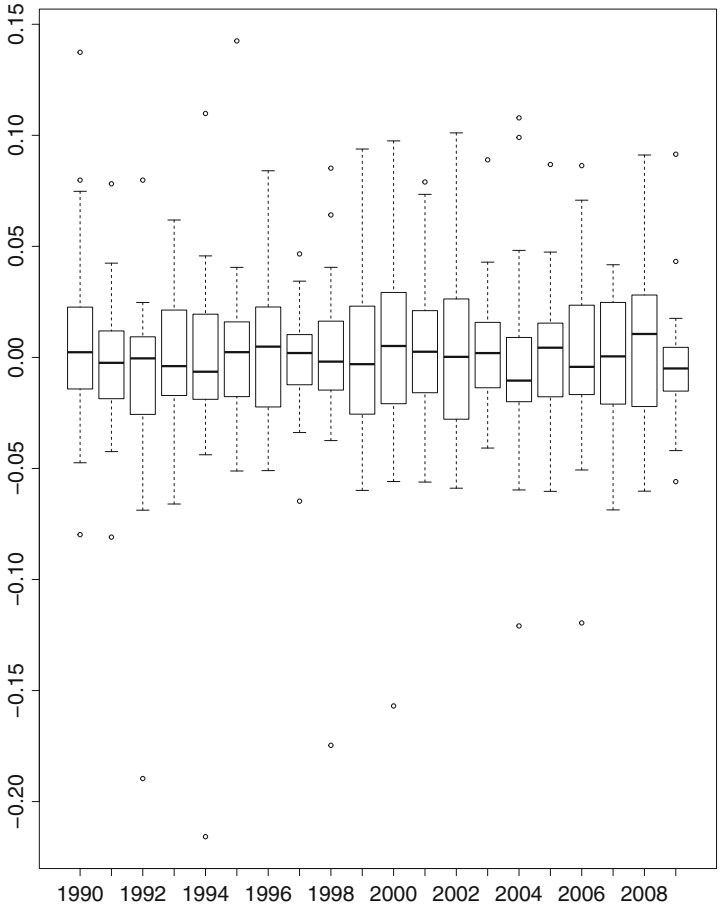


Fig. 6 Box plot of the residuals of SLMFE model for each year

Moreover, Slovenia is the only country in Eastern Europe that has an observed mortality higher than expected as a part of Europe. These results are confirmed in papers such as Trnka et al. (1998) and Zwerling et al. (2011). Those authors indicate that Slovenia was the only country in Eastern Europe, in which the revaccination and tuberculin skin tests were not applied. In eastern countries the prevalence of tuberculosis was very high so the mass primary vaccination and general revaccination (1994–1996) were very common.

There is a significant spatial correlation in the SMR of the 26 European countries as the Global Moran Index indicates.

Locally, the Local Moran Index has detected two significant clusters of European countries until the year 2002. A cluster of high SMR formed by Eastern European countries (Lithuania, Latvia, Estonia, Ukraine, Belarus, Slovakia, Hungary, and Poland) and another cluster of low SMR composed of western European countries (Spain, Italy, France, Switzerland, Germany, Luxembourg, and Belgium). The

countries with nonsignificant values of the local Moran Index do not belong to any cluster (The Czech Republic, The United Kingdom, Denmark, Finland, Ireland, The Netherlands, Norway, Slovenia, Portugal, Sweden, and Austria).

It is important to emphasize that the center (France) of cluster LL (low SMR) is unique and constant during the period 1990–2001. This cluster disappears from the year 2002 which indicates that the variability of the mortality in western countries has been increasing from 2002. On the contrary in the cluster HH (high SMR) several center clusters are observed because there are several clusters of type HH (high SMR) which form a unique macrocluster HH. These center clusters differ over the same period, moving from the west to east of Europe. In relation to the work of other authors, it is necessary to emphasize that to our knowledge, a spatial study to detect clusters of similar mortality in Europe, verifying in turn that the above-mentioned clusters are significant has not been carried out. There are some studies of mortality in Europe such as Meslé and Vallin (2002), Leon (2011), and Bonneux et al. (2010), but none of them take into account neighboring relations between countries to detect differences in mortality.

To model the detected space dependence in European countries, a Spatial Panel Data model was implemented. In particular, according to the typology of our data, a Spatial Lag Model with Fixed Effects (SLMFE). The dependent variable in the model is the logarithm of the SMR and the independent variables are: GDP, activity rate, and road sector energy consumption. The main conclusions of the implemented model are detailed.

The estimate for the spatial parameter (λ) is positive and highly statistically significant. This indicates that the logarithm of the SMR of a country varies with the logarithm of the SMR between its geographical neighbors. Specifically, the value of logarithm of the SMR of a country will increase by 34.65% when the logarithm of the SMR values corresponding to the environment also increases.

All covariates are significant which means that variations of three covariates cause variations in the logarithm of the SMR of a country.

The countries that form the different clusters are confirmed in the estimates of the spatial fixed effects of Spatial Lag Model. The estimations of spatial fixed effects with a negative sign belong to the cluster of western countries (low SMR) while estimations of spatial fixed effects with a positive sign belong to the cluster of eastern countries (high SMR). Most countries with non-significant spatial effects do not form a cluster.

As regards time effects, all of them are not significant. To study whether the fixed effects should be excluded from the model the Lagrange Multiplier test was applied. The test concluded that neither the spatial nor time effect can be excluded from the model. In addition, time effects were represented graphically which follow a growing trend except for the period 1994–1998. This unfavorable evolution picks up the collapse of the Soviet system.

Finally, the SLMFE is a good model as the measures of goodness of fit indicate (the value of residual variance is 0.001 and the value of the determination coefficient is 0.97). The analysis of the residuals of the SLMFE model shows that the model takes into account the spatial dependence over the period 1990–2009.

There is literature showing that western European countries have a higher life expectancy than the rest, such as Mackenbach et al. (2013), Vågerö (2010), and Meslé (2004). Many of these studies address the well-known life expectancy, but none uses a spatial methodology to detect significant associations between countries with similar mortality implementing in turn a spatial model which controls the space dependence of these countries over time. These spatial panel data models, some of which are still in their early development, are applied in fields as varied as sociology, epidemiology, geology, criminology, etc. (Gersmehl, 2014), but have not been implemented in the actuarial field.

Acknowledgements Support for the research presented in this paper was provided by a grant from Ministry of Economy and Competitiveness, project MTM2013–45381-P.

References

Anselin, L.: Local indicators of spatial association–LISA. Geogr. Anal. **27**(2), 93–115 (1995)
Arbia, G., Piras, G.: Convergence in per-capita GDP across European regions using panel data models extended to spatial autocorrelation effects. ISAE Working Paper No. 51. Available at SSRN: https://ssrn.com/abstract=936327 (2005) or doi:http://dx.doi.org/10.2139/ssrn.936327
Asteriou, D., Hall, S.: Applied Econometrics. Palgrave Macmillan, New York (2015)
Baltagi, B.: Econometric Analysis of Panel Data. Wiley, New York (2008)
Bivand, R., Lewin-Koh, N.: maptools: tools for reading and handling spatial objects. R package version 0.8–29 (2014)
Bivand, R., Keitt, T., Rowlingson, B.: rgdal: Bindings for the Geospatial Data Abstraction Library. R Package Version 1.1-10. https://CRAN.R-project.org/package=rgdal (2016)
Bivand, R.: spdep: Spatial Dependence: Weighting Schemes, Statistics and Models. R Package Version 0.5-53. http://CRAN.Rproject.org/package=spdep (2012)
Bobadilla, J.L., Costello, C.A., Mitchell, F., et al.: Premature Death in the New Independent States, pp. 61–239. National Academies Press, Washington, DC (1997)
Bonneux, L., Huisman, C., de Beer, J.: Mortality in 272 European regions, 2002–2004: an update. Eur. J. Epidemiol. **25**(1), 77–85 (2010)
Breusch, T.S., Pagan, A.R.: The Lagrange multiplier test and its applications to model specification in econometrics. Rev. Econ. Stud. **47**(1), 239–253 (1980)
Charpentier, A.: Spatial analysis. In: Computational Actuarial Science with R. Chapman and Hall/CRC, Boca Raton (2014). ISBN: 978-1-4665-9259-9
Croissant, Y., Millo, G., et al.: Panel data econometrics in R: the plm package. J. Stat. Softw. **27**(2), 1–43 (2008)
Cutler, D., Deaton, A., Lleras-Muney, A.: The determinants of mortality. J. Econ. Perspect. **20**(3), 97–120 (2006)
Diehl, P.: Geography and war: a review and assessment of the empirical literature. In: Ward, M. (ed.) The New Geupolitics, pp. 37–121. Gordon and Breach, Philadelphia, PA (1992)
Elhorst, J.P.: Spatial Econometrics: From Cross-Sectional Data to Spatial Panels. Springer, New York (2014a)
Elhorst, J.P.: Spatial panel models. In: Handbook of Regional Science, pp. 1637–1652. Springer, New York (2014b)
EUROSTAT: Health Statistics-Atlas on Mortality in the European Union. European Communities, Luxembourg (2009)
EUROSTAT: Quality of life indicators. http://ec.europa.eu/eurostat/statistics-explained/index.php/Quality_of_life_indicators_health (2013)

Fleiss, J., Levin, B., Paik, M.: Statistical Methods for Rates and Proportions. Wiley Series in Probability and Statistics. Wiley, Hoboken (2013)
Gersmehl, P.: Teaching Geography, 3rd edn. Guilford, New York (2014)
Gordon, M.: Gmisc: descriptive statistics, transition plots, and more. R package version 1.3.1 (2016)
Hinde, A.: Demographic Methods. Routledge, London (1998)
Hsiao, C.: Analysis of Panel Data, 3rd edn., p. 38. Cambridge University Press, Cambridge (2014)
Hsiao, C., Pesaran, M.H., Tahmiscioglu, A.K.: Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. J. Econometr. **109**(1), 107–150 (20022)
Human Mortality Database: University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Avaliable at www.mortality.org or www.humanmortality.de (2014). Accessed 17 Apr 2014
Hyndman, R.J., Booth, H., Tickle, L., Maindonald, J.: Demography: forecasting mortality, fertility, migration and population data. R package version 1.18 (2014)
Kennedy, P.: A Guide to Econometrics. MIT Press, Cambridge (2003)
Laurent, T., Ruiz-Gazen, A., Thomas-Agnan, C.: GeoXp: an R package for exploratory spatial data analysis. J. Stat. Softw. 47(2):1–23 (2012)
Leon, D.A.: Trends in European life expectancy: a salutary view. Int. J. Epidemiol. **40**, 271–277 (2011)
Mackenbach, J.P., Karanikolos, M., McKee, M.: The unequal health of Europeans: successes and failures of policies. Lancet **381**(9872), 1125–1134 (2013)
Meslé, F.: Mortality in Central and Eastern Europe: long-term trends and recent upturns. Demogr. Res. **S2**:45–70 (2004)
Meslé, F., Vallin, J.: Mortality in Europe: the divergence between East and West. Population (English Edition) **57**(1), 157–197 (2002)
Millo, G., Piras, G.: splm: Spatial panel data models in R. J. Stat. Softw. **47**(1):1–38 (2012)
Moran, P.: Notes on continuous stochastic phenomena. Biometrika **37**(1–2), 17–23 (1950a)
Moran, P.: A test for the serial independence of residuals. Biometrika **37**(1–2), 178–181 (1950b)
Neuwirth, E.: RColorBrewer: ColorBrewer palettes. R package version 1.1–2 (2014)
R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/ (2015)
Spinakis, A., Anastasiou, G., Panousis, V., Spiliopoulos, K., Palaiologou, S., Yfantopoulos, J.: Expert Review and Proposals for Measurement of Health Inequalities in the European Union. European Commission. Technical report, European Commission Directorate General for Health and Consumers, Luxembourg (2015). ISBN 978-92-79-18529-8
The World Bank Database: World development indicators. Data retrieved from World Development Indicators, http://data.worldbank.org/ (2015). Accessed Jan 2015
Torgo, L.: Data Mining with R, Learning with Case Studies. Chapman and Hall/CRC, Boca Raton (2010)
Trnka, L., Dankova, D., Zitova, J., Cimprichova, L., Migliori, G.B., Clancy, L., Zellweger, J.: Survey of BCG vaccination policy in Europe: 1994–96. Bull. World Health Organ. **76**(1), 85–91 (1998)
Vågerö, D.: The East-West health divide in Europe: growing and shifting eastwards. Eur. Rev. **18**(1), 23–34 (2010)
Vaupel, J.W., Zhang, Z., van Raalte, A.A., Vaupel, J.W., Zhang, Z., van Raalte, A.A.: Life expectancy and disparity: an international comparison of life table data. BMJ Open, 1(1), 1–6 (2011)
Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data MIT Press, Cambridge (2010)
Zwerling, A., Behr, M.A., Verma, A., Brewer, T.F., Menzies, D., Pai, M.: The BCG World Atlas: a database of global BCG vaccination policies and practices. PLoS Med. **8**(3), e1001012 (2011)

Stochastic Control for Insurance: Models, Strategies, and Numerics

Christian Hipp

Abstract This survey on stochastic control for insurance is written for stimulation research of the topic, addressing new problems (such as dividend values with ruin constraint) and new methods (as the non-stationary approach) as well as numerical issues (Euler type discretizations). In the context of discretizations, viscosity arguments are important which are adapted here for the purpose of solving insurance problems. Finally, open problems are listed.

Keywords Stochastic control • Viscosity solutions • Euler type discretizations • Multi objective problems

AMS classification: primary 91B30, 93E20; secondary 49l20, 49L25, 49M25

1 Prologue

This paper is based on a short course given at University of Cartagena, Columbia, during the Second International Congress on Actuarial Science and Quantitative Finance. Its issue is an introduction into stochastic control in insurance, with special emphasis on new problems, new approaches and new methods, as well as on numerical issues. We will consider control for minimizing ruin probability (which results in reduction of solvency capital) as well as maximizing dividend payment (which has impact on the company value). Combining these two objectives, we consider maximization of dividend value under a ruin constraint. We will start with a simple discrete example where the tools and methods for more complex models are introduced. This example is just for illustration, it is too simple for applications or for advanced mathematics. Such simple models have their merits in education (see, e.g., De Finetti 1957). In this discrete example we consider

C. Hipp (Retired)
Karlsruhe Institute of Technology, Institute for Finance and Insurance, Karlsruhe, Germany
e-mail: FChristian.Hipp@gmail.com

J.A. Londoño et al. (eds.), *Actuarial Sciences and Quantitative Finance*, Springer
Proceedings in Mathematics & Statistics 214, https://doi.org/10.1007/978-3-319-66536-8_4

1. infinite time ruin probability,
2. minimal ruin probability by control of reinsurance,
3. minimal ruin probability by control of investment,
4. company value, i.e. maximal dividend value by control of dividend payment,
5. maximal company value by control of reinsurance,
6. company value under ruin constraint, and
7. maximal company value under ruin constraint with control of reinsurance.

A Discrete Example We consider a discrete time and space risk process $S(t), t \geq 0$, which jumps from s to $s+2$ with probability $p_1 = 0.55$, to $s-1$ with probability $p_2 = 0.3$, and to $s-3$ with probability $p_3 = 0.15$. This can be regarded as a risk process of an insurer who in each period receives a premium of size 2 and pays claims of size 3 and 5, respectively. The infinite horizon **ruin probability**

$$\psi(s) = \mathbb{P}\{S(t) < 0 \text{ for some } t \geq 0 | S(0) = s\}$$

satisfies the dynamic equation

$$\psi(s) = p_1\psi(s+2) + p_2\psi(s-1) + p_3\psi(s-3), s \geq 0, \tag{1}$$

with $\psi(s) = 1$ for $s < 0$. Using the operator

$$\mathcal{G}f(s) = p_1 f(s+2) + p_2 f(s-1) + p_3 f(s-3)$$

the above dynamic equation reads

$$\psi(s) = \mathcal{G}\psi(s), s \geq 0.$$

The common computation of $\psi(s)$ is done via generating functions, the solution of the characteristic equation and the adjustment to the boundary values $\psi(s) = 1, s < 0$ and $\psi(\infty) = 0$. The characteristic equation

$$z^3 = p_1 z^5 + p_2 z^2 + p_3$$

has the five complex solutions z_i which with coefficients C_i form the solution $\psi(s) = C_1 z_1 + \ldots + C_5 z_5$ having the appropriate boundary values. In our example, in particular, $\psi(12) = 0.08828824$.

i	z_i	C_i
1	1	0
2	0.835935	0.758246
3	-1.503707	0
4	$-0.166114 + 0.435170i$	$-0.006270 + 0.020799i$
5	$-0.166114 - 0.435170i$	$-0.006270 - 0.020799i$

For numerical computation and for the next problems it is useful to consider instead a nonstationary approach. For $t \geq 0$ define $\psi(s,t)$ as the probability of ruin after time t, given $S(t) = s$. The functions $s \to \psi(s,t)$ satisfy

$$\psi(s,t-1) = \mathcal{G}\psi(s,t), s \geq 0, \tag{2}$$

with $\psi(s,t) = 1$ for $s < 0$. Starting with large $T > 0$ and initial function $\phi(s,T) = 0, s \geq 0$, $\phi(s,T) = 1, s < 0$ we calculate with (2) for the functions $\phi(s,t)$ all terms down to $t = 0$, and $\phi(s,0)$ is a good approximation for $\psi(s) : \phi(s,0)$ is the probability for ruin before or at T which is close to $\psi(s)$ when T is large. For $T = 5000$ we obtain for $\phi(12,0)$ all the digits for $\psi(12)$ shown above.

Assume that for each period we can buy reinsurance: for the price of 1 the reinsurer pays 3 when a claim of size 5 occurs, and 1 when the claim has size 3. So for each claim the first insurer has to pay 2; this type of risk sharing is called excess of loss reinsurance. What is the **optimal reinsurance** strategy to **minimize the ruin probability**, and what is the corresponding ruin probability $\overline{\psi}(s)$? The nonstationary approach—with a slightly changed dynamic equation—produces the solution: replace (2) by

$$\overline{\psi}(s,t-1) = \min[\mathcal{G}\overline{\psi}(s,t), \mathcal{G}_1\overline{\psi}(s,t)] \tag{3}$$

$$\mathcal{G}_1 f(s) = p_1 f(s+1) + p_2 f(s-1) + p_3 f(s-1)]. \tag{4}$$

The operator $\mathcal{G}$ shows the dynamics in the case without reinsurance, while the operator in (4) corresponds to the dynamics with reinsurance. The numerical procedure and the initial functions are the same as above. With dynamic reinsurance, the ruin probability is reduced to $\overline{\psi}(12) = 0.063095$. The optimal reinsurance strategy is: buy reinsurance whenever $s \geq 2$. With static reinsurance, i.e. reinsurance for all $s \geq 0$, we obtain $\overline{\psi}(12) = 0.073629$.

Assume that for each period we can invest an amount of 1 which in this period either doubles with probability $w > 1/2$, or is lost with probability $1 - w$, where investment return is independent of insurance business. What is the **optimal investment** strategy to minimize the ruin probability? The nonstationary approach—again with a slightly changed dynamic equation—produces the solution: replace (2) by

$$\overline{\psi}(s,t-1) = \min[\mathcal{G}\overline{\psi}(s,t), \mathcal{G}_2\overline{\psi}(s,t)] \tag{5}$$

$$\mathcal{G}_2 f(s) = w\mathcal{G}f(s+1) + (1-w)\mathcal{G}f(s-1) \tag{6}$$

The operator $\mathcal{G}$ shows the dynamics in the case without investment, while the operator in (6) corresponds to the dynamics with investment. The numerical procedure and the initial functions are the same as above. With dynamic investment, the ruin probability for $w = 0.55$ is reduced to $\tilde{\psi}(12) = 0.07611$. The optimal investment strategy is: invest whenever $s \notin \{0,2\}$. With static investment, i.e.

investment for all $s \geq 0$, we obtain $\tilde{\psi}(12) = 0.0923987$ which is larger than without investment. Of course, one might extend the control to invest more than 1, which can be solved with a larger set of operators.

For discount rate $0 < r < 1$ and for a given dividend strategy $d(t), t \geq 0$, we consider the expected discounted dividends

$$V^d(s) = E\left[\sum_{n=0}^{\infty} r^n d(n) | S(0) = s\right],$$

where $d(t)$ is paid at time t and depends on the history up to time $t-1$. The dividend risk process is $S^d(s) = S(t) - d(0) - \ldots - d(t-1)$; its ruin time is denoted by τ^d. Dividend payments are forbidden at and after ruin. The **company value** is given by

$$V(s) = \sup_d V^d(s), s \geq 0,$$

its dynamic equation equals

$$V(s) = \max\{r\mathcal{G}V(s), V(s-1) + 1\}, \tag{7}$$

with $V(s) = 0$ for $s < 0$. As above, the generating function method can be applied here, but this equation can also be solved with a nonstationary approach. For $t \geq 0$ consider the time t dividend functions

$$V^d(s,t) = E\left[\sum_{n=t}^{\infty} r^n d(n) | S(t) = s\right],$$

and define $V(s,t)$ as the supremum of these dividend functions. The functions $V(s,t)$ satisfy

$$V(s,t-1) = \max\{\mathcal{G}V(s,t), V(s-1,t-1) + r^{t-1}\}, \tag{8}$$

with $V(s,t) = 0,\ s < 0$. Starting with large $T > 0$ and initial function $V(s,T) = 0$ we calculate with (8) the functions $V(s,t)$ down to $t = 0$, and $V(s,0)$ is a good approximation for $V(s)$. For $r = 0.98$ we obtain $V(12) = 17.933928$.

For simultaneous **control of dividend payments and reinsurance** we simply replace the expression $\mathcal{G}V(s,t)$ in (8) by $\max(\mathcal{G}V(s,t), \mathcal{G}_1 V(s,t))$. The dividend value changes to $V(12) = 18.104876$. The optimal reinsurance strategy is: buy reinsurance when $s \geq 10$.

Optimal **dividend payment with a ruin constraint** has the value function

$$V(s,\alpha) = \sup_d [V^d(s) : \mathbb{P}\{\tau^d < \infty\} \leq \alpha].$$

To solve it we use the Lagrange multiplier method and maximize for a large constant L the expression

$$W(s,L) = \sup_d [V^d(s) - L\mathbb{P}\{\tau^d < \infty\}].$$

For the nonstationary approach we consider again the dividend payments after time t, together with the ruin time τ_t^d after time t :

$$W(s,t) = \sup_d [V^d(s,t) - L\mathbb{P}\{\tau_t^d < \infty | S(t) = s\}].$$

The dynamic equation for these functions reads

$$W(s,t-1) = \max\{\mathcal{G}W(s,t), W(s-1,t-1) + r^{t-1}\}, \tag{9}$$

where $W(s,t) = -L$ for $s < 0$. The initial function here is $W(s,T) = -L\psi(s)$. For the computation of the corresponding ruin probability, we simultaneously compute functions $\psi(s,t)$ from

$$\psi(s,t-1) = \mathcal{G}\psi(s,t), \tag{10}$$

when the maximum in (9) is at $\mathcal{G}W(s,t)$ (no dividend payment), and $\psi(s,t-1) = \psi(s-1,t-1)$ otherwise. The initial function is $\psi(s,T) = \psi(s)$. For $s = 12$ we have a ruin probability without dividend payment $\psi(12) = 0.088288$ and a dividend value without constraint $V(12) = 18.933928$. We take $L = 40$ and obtain $W(12) = 5.646781$. The ruin probability with dividend payments equals $\psi(12) = 0.160923$, so the dividend value is $W(12) + L\psi(12) = 12.083708$.

For simultaneous **control of dividend payments under a ruin constraint and reinsurance** we simply replace the expression $\mathcal{G}W(s,t)$ in (9) by

$$\max(\mathcal{G}W(s,t), \mathcal{G}_1 W(s,t)).$$

Also here, we obtain the corresponding ruin probability in a simultaneous computation: we use the dynamic equations (10) when no dividends are paid and no reinsurance is bought, or the relation

$$\psi(s,t-1) = \mathcal{G}_1\psi(s,t)$$

when no dividends are paid and reinsurance is bought, or finally

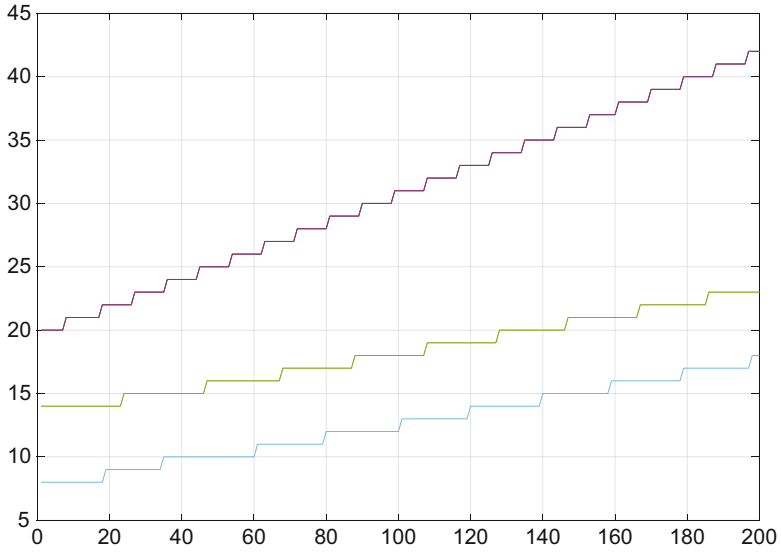
$$\psi(s,t-1) = \psi(s-1,t-1)$$

when dividends are paid. For $s = 12$ and $L = 13.754$ we obtain a ruin probability $\psi(12) = 0.160828$ and a dividend value $W(12,L) + L\psi(12) = 17.635244$. Comparing this company value with the one without ruin constraint, one can see that

a ruin constraint can be cheap when an appropriate reinsurance cover is available (which, in our case, has a rather small loading: the premium is 1 while the expected payments are 0.75).

Apparently, the nonstationary approach seems to be well suited for the computation of value functions and optimal strategies since it can easily be adapted to various different problems. It is superior to the stationary method given in Hipp (2003) which is based on a modified Hamilton-Jacobi-Bellman equation: it is much faster. This is caused by the fact that in the stationary approach one has to compute value functions for all $0 \le \alpha \le 1$, while in the nonstationary approach one has only one fixed α (specified by L).

The following figure shows the optimal strategies for control of dividends with ruin constraint with and without reinsurance. They depend on the current surplus s and time t. In both cases the optimal dividend strategies are barrier strategies defined by a barrier $M(t)$; the optimal reinsurance strategy is also a barrier strategy: buy reinsurance when $s \ge N(t)$. The values of $M(t)$ and $N(t)$ are piecewise constant; they are shown for $t = 0, \ldots 200$. The highest curve shows $M(t)$ for the case without reinsurance, while the two curves below show $M(t)$ and $N(t)$ for the case with reinsurance.



Continuous Time Models and Their Generators For applications, continuous time models are of major importance. The classical risk model for insurance is the **Lundberg model** in which the claim arrivals are modeled as a homogeneous Poisson process $N(t), t \ge 0$, with constant intensity $\lambda > 0$, and the claim sizes $X, X_1, X_2, \ldots$ are independent and identically distributed and independent of the process $N(t)$. The risk process is then given by

$$S(t) = s + ct - X_1 - \ldots - X_{N(t)},$$

where s is the initial surplus and c the premium rate. We always assume a positive loading, i.e. $c > \lambda E[X]$. $S(t)$ is a time homogeneous process with independent increments. The above operator $\mathcal{G}$ in continuous time homogeneous Markov processes is the infinitesimal generator

$$\mathcal{G}f(s) = \lim_{h \to 0} E[f(S(h) - f(s)|S(0) = s]/h, \tag{11}$$

which for the Lundberg model equals

$$\mathcal{G}f(s) = \lambda E[f(s - X) - f(s)] + cf'(s)$$

which is defined on the set of all bounded differentiable functions $f(s)$.

A large scale approximation of stationary risk processes with independent increments is the **simple diffusion** with dynamics

$$dS(t) = \mu dt + \sigma dW(t),\ t \geq 0, \tag{12}$$

where $W(t)$ is the standard Brownian motion. The generator equals

$$\mathcal{G}f(s) = \mu f'(s) + \sigma^2 f''(s)/2,$$

it is defined on the set of locally bounded functions with second derivative.

One possible way to include parameter uncertainty is the choice of mixture models for $S(t)$, such as the **Cox process** in which the intensity of the claims arrival process is random and modeled as a time homogeneous finite Markov process. Here, we have a finite number of possible non-negative and distinct intensities $\lambda_i,\ i = 1, \ldots, m$, and $\lambda(t)$ jumps between these intensities in a homogeneous Markovian way. This is usually described via parameters $b_{i,j},\ i, j = 1, \ldots, m$, satisfying $b_{i,j} \geq 0,\ i \neq j$, and

$$b_{i,i} = -\sum_{j \neq i} b_{i,j}.$$

If the intensity is in state λ_i, then it stays there an exponential waiting time with parameter $-b_{i,i}$, and then it jumps to λ_j with probability $-b_{i,j}/b_{i,i}$.

Mixture models are more complex than the above-mentioned models, they sometimes lack the independence of increments and often also the Markov property. When $\lambda(t)$ is observable, then the state $(S(t), \lambda(t))$ has the Markov property, and the generator for this at $s \geq 0,\ i = 1, \ldots, m$ equals

$$\mathcal{G}f(s, i) = \lambda_i E[f(s - X, i) - f(s, i)] + cf_s(s, i) + \sum_{j=1}^{m} b_{i,j}(f(s, j) - f(s, i)). \tag{13}$$

When $\lambda(t)$ is not observable, then again one can enlarge the state vector to obtain the Markov property: if $\mathcal{F}_t$ is the filtration generated by $S(t),\ t \geq 0$, then $(S(t), p_1(t), \ldots, p_m(t))$ has the Markov property, where $p_i(t)$ is the conditional probability of $\lambda(t) = \lambda_i$, given $\mathcal{F}(t)$. The processes $p_k(t)$ are piecewise deterministic, they depend on t and the history $S_u, u \leq t$. Between claims they can be computed using the following system of interacting differential equations:

$$p_k'(t) = \sum_{j=1}^{I} p_j(t) b_{j,k} - \lambda_k p_k(t) + p_k(t) \sum_{j=1}^{I} p_j(t)\lambda_j,\ i = 1, \ldots, I. \tag{14}$$

This follows from the fact that from t to $t+dt$, given $\lambda(t) = \lambda_k$, there is no transition and no claim with probability $1-\lambda_k dt+b_{k,k}dt+o(dt)$, and for $j \neq k$, given $\lambda(t) = \lambda_j$, there is a transition from λ_j to λ_k and no claim with probability $b_{j,k}dt + o(dt)$. So, given $N(t + dt) = N(t)$,

$$\begin{aligned} p_k(t + dt) &= \frac{p_k(t)\,(1 - \lambda_k dt + b_{k,k}dt) + \sum_{j\neq k} b_{j,k} p_j(t)dt}{P\{N(t + dt) = N(t)|\mathcal{F}_t\}} + o(dt) \\ &= \frac{p_k(t)\,(1 - \lambda_k dt) + \sum_j b_{j,k} p_j(t)dt}{1 - \sum_j p_j(t)\lambda_j dt} + o(dt) \\ &= p_k(t)\left(1 - \lambda_k dt + \sum_j p_j(t)\lambda_j dt\right) + \sum_j b_{j,k} p_j(t)dt + o(dt). \end{aligned}$$

At a claim, the process $p_k(t)$ has a jump: given $N(t + dt) > N(t)$ we have for $k = 1, \ldots, I$

$$p_k^+ := p_k(t+) = \frac{\lambda_k p_k(t)}{\sum_j p_j(t)\lambda_j}. \tag{15}$$

This follows from

$$\begin{aligned} P\{N(t + h) > N(t), \lambda(t + h) = \lambda_k|\mathcal{F}_t\} &= p_k(t)\lambda_k h + o(h), \\ P\{N(t + h) > N(t)|\mathcal{F}_t\} &= \sum_{j=1}^{I} p_j(t)\lambda_j h + o(h). \end{aligned}$$

From this dynamics we obtain the following generator:

$$\mathcal{G}f(s, p) = \sum_k p_k \lambda_k E[f(s - X, p^+) - f(s, p)] + cf_s(s, p) + \sum_k f_{p_k}(s, p)p_k'. \tag{16}$$

Here, f_s and f_{p_k} are the partial derivatives with respect to s and p_k, respectively.

Mixtures with constant but unknown parameters in a finite set $\{\lambda_1, \ldots, \lambda_m\}$ can be modeled as follows: let λ be a random variable with $p_i = \mathbb{P}\{\lambda = \lambda_i\}$ known. Assume that given $\lambda = \lambda_i$, $S(t)$ is a classical Lundberg process with intensity λ_i. With $p_i(t)$ the conditional probability of $\lambda = \lambda_i$, given $S(u)$, $u \leq t$, the vector $(S(t), p_1(t), \ldots, p_m(t))$ has the Markov property. The dynamics of the $p_i(t)$ is the same as in the above example, with $b_{j,k} = 0$ for $j, k = 1, \ldots, m$. The generator is the same as in (16).

For mixture models as well as for dividend problems with ruin constraint, it is convenient to consider also non-stationary generators. As illustration we mention the example which is also considered in Sect. 5. It is a delayed compound Poisson process where up to a random time T we have $S(t) = s + ct$, and for $t > T$, given T the risk process $s + S(t) - S(T)$ is a compound Poisson process. The time T has an exponential distribution. We want to minimize the ruin probability by control of reinsurance. For this, write $V(s, t)$ for the controlled ruin probability after time t, given that no claim happened until t. Then $V(s, t)$ has a dynamic equation of the form

$$0 = \inf_a p(t)\lambda E[V_1(s - g_a(X)) - V(s, t)] + (c - h(a))V_s(s, t) + V_t(s, t),$$

where $p(t)$ is the conditional probability of $t < T$, given no claim up to time t, and $V_1(s)$ is the minimal ruin probability for the case with constant positive intensity. The quantity $h(a)$ is the reinsurance price for risk sharing $g_a(X)$.

2 Ruin and Company Value

We shall restrict ourselves to three types of control problem: one in which we minimize the infinite time of ruin, next the maximization of the company value, and finally the maximization of a company value with a ruin constraint. We shall always consider an infinite horizon view, since insurance uses diversification in time, and some insurance products are long term.

For Lundberg models, the infinite time **ruin probability** is a classical bounded solution of the dynamic equation $\psi(s) = \mathcal{G}\psi(s), s \geq 0$, with a continuous first derivative. It is the unique classical solution satisfying $\psi(s) = 1$, $s < 0$, $\psi(\infty) = 0$, and $\psi'(0) = -\lambda(1 - V(0))/c$. Analytic expressions for $\psi(s)$ can be given for exponential or more general phase-type distributions (see Chap. IX of Albrecher and Asmussen 2010).

The **company value** is itself the result of a control problem: what is the maximal expected discounted sum of paid dividends? In mathematical terms:

$$V_0(s) = \sup_D \left\{ E\left[\int_0^\infty e^{-\delta t} dD(t) | S(0) = s \right] \right\},$$

where $D = D(t),\ t \geq 0$ is the sum of dividends paid up to time t with some admissible dividend payment strategy. Already in the Lundberg model, this question is hard, too hard for applications in insurance. The answer is simpler if we restrict dividend payment to strategies which are barrier strategies. Optimal barrier strategies can be derived from a classical solution $v(s)$ of the dynamic equation

$$0 = \delta v(s) + \mathcal{G}v(s), \tag{17}$$

with $v(0) = v'(0) = 1$, where $\mathcal{G}$ is the generator of the risk process and δ is the discount rate. Then

$$V_0(s) = v(s)/v'(M),\ s \leq M,\ V_0(s) = V_0(M) + s - M,\ s \geq M,$$

where the optimal barrier is given by

$$M = \arg\min v'(s)$$

(see Schmidli 2007, Sect. 2.4.2). This simplified answer is a sub-solution of the above control problem. It is an optimal dividend strategy only for special claim size distributions (see Loeffen 2008). Generally, optimal dividend strategies are band strategies, i.e. there might exist $M < M_1 < M_2$ for which no dividends are paid as long as $M_1 < S(t) < M_2$, and for $M < S(t) \leq M_1$ a lump sum $M_1 - S(t)$ is paid out immediately. However, optimal barrier strategies are useful for applications since for $s \leq M$ the dividend values of the barrier strategy are the same as the dividend value of the optimal band strategy.

For the company value, a discount rate is needed which can be a market interest rate (which should be modeled with some stochastic process which is allowed to be negative), or a value which shareholders and accountants agree upon. We will be concerned only with positive constant discounting rates.

A **company value with ruin constraint** is an even more complex quantity since it involves a control problem with two objective functions. Its computation is still work in progress. We consider it here since it appealing to both, the policy holders and the stock holders. The value is given by

$$V(s,\alpha) = \sup_D \left\{ E\left[\int_0^\infty e^{-\delta t} dD(t) | S(0) = s \right] : \psi^D(s) \leq \alpha \right\},$$

where $0 < \alpha \leq 1$ is the allowed ruin probability and $\psi^D(s)$ is the with dividend ruin probability. Clearly, $V(s,1) = V_0(s)$, and if $\psi(s)$ is the without dividend ruin probability, then $V(s,\psi(s)) = 0$.

The meaning of a company value with ruin constraint might become clearer when we meditate a little about special dividend strategies which have constrained ruin probabilities. Let us do this in a diffusion model which does not have downward jumps. Let $s(\alpha)$ be the solution of $\psi(s) = \alpha$. The simplest strategy is: pay out $s - s(\alpha)$ immediately, and stop dividends forever. This has a ruin probability α

and a dividend value $s - s(\alpha)$. A better strategy is constructed using the optimal unconstrained dividend strategy based on the barrier M and leading to the dividend value $V_0(s)$. Choose $s > 0$ and $\alpha > \psi(s)$; then $s(\alpha) < s$, so you can put aside $s(\alpha)$ (e.g., into your pocket), and then use the unconstrained dividend strategy with initial surplus $s-s(\alpha)$. At ruin, i.e. when you reach zero, you stop paying dividends forever. With your money from the pocket, you indeed stopped with $s(\alpha)$, and so your ruin probability equals α. And your corresponding dividend value equals $V_0(s-s(\alpha)) > s - s(\alpha)$.

Money in the pocket is never optimal, and so there should exist improvements of the dividend strategy with the same ruin probability. Our next strategy is based on the improvement procedure introduced in Hipp (2016). We assume $s < M$ and $\alpha > \psi(s)$. Do not pay dividends until you reach M. You will be ruined before reaching M with probability $A = (\psi(s) - \psi(M))/(1 - \psi(M)) \leq \psi(s) < \alpha$. Define $0 < \gamma < \alpha$ via equation

$$A + \gamma(1 - A) = \alpha.$$

When you reach M, you put aside the amount $s(\gamma)$ and pay out dividends with the unconstrained strategy until you reach $s(\gamma)$. Then you again stop paying dividends forever. The resulting ruin probability is α, and the dividend value will be $V_0(M - s(\gamma))$, discounted over the time τ until you reach M. With our function $V_0(s)$ above we have $E[e^{-\tau\delta}] = V_0(s)/V_0(M)$, and so the dividend value of our strategy is

$$\frac{V_0(s)}{V_0(M)} V_0(M - s(\gamma))$$

which is larger than $V_0(s - s(\alpha))$. The reason for this is: in the first case we stop dividend payment forever at $s(\alpha)$, also when we did not reach M yet, and this reduces the dividend payments. In the second we wait until we reach M, and then money goes to our pocket.

3 Hamilton-Jacobi-Bellman Equations

The use of these equations might seem a bit old-fashioned, but with the concept of viscosity solutions it is still a standard approach. For a stationary Markov process which should be controlled by actions $a \in A$ we consider the process with a constant (in time) action a and the corresponding generator $\mathcal{G}^a$ of the resulting Markov process. If we minimize ruin probability, the Hamilton-Jacobi-Bellman equation reads

$$0 = \inf_a \mathcal{G}^a V(x), \tag{18}$$

where $V(x)$ stands for the value function of the problem and $x = (s, p)$ is the (enlarged) vector of states, $s \geq 0$ being the surplus. If we maximize dividend payments, it is given by the formula

$$0 = -\delta V(s) + \sup_a \mathcal{G}^a V(x), \tag{19}$$

where $\delta > 0$ is the discount rate. But here the range of x is restricted to $\{(x, p) : x \leq M(p)\}$, where for fixed p, $M(p)$ is the smallest point x at which $V_s(x, p) = 1$. For larger x the function is linear with slope 1:

$$V(x, p) = V(M(p), p) + x - M(p).$$

Notice that we neglect a possible second and third, etc. band.

In Lundberg models, Eq. (18) involves a first derivative and an expectation. Such an equation needs two boundary values to identify a unique solution. For ruin probabilities $V(s) = 1$ for $s < 0$, and so we can use the two conditions $V(\infty) = 0$ and $V'(0) = \lambda(1-V(0)/c$. For dividend values we first use a solution with $v(s) = 0$ for $s < 0$ and $v(0) = 1, v'(0) = \lambda/c$, and then we minimize $v'(s)$ (see Chap. 6).

In simple diffusion models, Eq. (18) shows a first and a second derivative. For this we again need two conditions, which are $V(0) = 1, V(\infty) = 0$ for the ruin probability, and $V(0) = 0, V'(M) = 1$ for dividend values, where M is again the minimizer for $v'(s)$.

We shall frequently use a nonstationary approach, even for stationary problems. In our introductory example, we have computed the infinite horizon ruin probability with such an approach: we considered the ruin probability $V(s, t)$ after time t when starting in s at t. We used a large number T and used the initial guess $V(s, T) = 1$ if $s < 0$, and $V(s, T) = 0$ elsewhere. Using the dynamic equation for the nonstationary case, we calculated backward in t to the end $t = 0$, and $V(s, 0)$ was an almost exact value for the ruin probability in the stationary model.

For this we need the dynamic equation for a nonstationary setup in the case of a stationary Markov model. This is most simple: if $\mathcal{G}$ is the generator of the model, then the equation is

$$0 = \mathcal{G}V(s, t) + V_t(s, t). \tag{20}$$

In the dividend case, there is no extra term with δ as in (19) since the discounting is modeled in the time dependence.

For cases like the volcano problem in Chap. 5, we obtain a nonstationary dynamic equation in which time dependent quantities enter.

4 Investment Control

What is the optimal investment strategy for an insurer to minimize her ruin probability? This is one of the oldest questions in the field of stochastic control in insurance, it was solved for the simple diffusion case by Browne (1995) in 1995. A simple framework for this problem is a Lundberg process for the risk and a logarithmic Brownian motion for the capital market (a stock or an index) in which the insurer can invest.

Our first example is of little use in insurance industry, but it might serve as an introduction since it shows many features which are present in other cases. We assume that the insurer does not earn interest and pay taxes, and that he can invest an unrestricted amount, i.e. unlimited leverage and short-selling are allowed. We assume in the following that the Lundberg process has parameters c (premium rate), λ (claim frequency), X (generic claim size) with bounded density, and $c > \lambda E[X]$ (positive loading).

The price process of the asset has dynamics

$$dZ(t) = \mu Z(t)dt + \sigma Z(t)dW(t),\ t \geq 0,$$

where $W(t)$ is standard Brownian motion and μ, σ are positive.

Theorem 1 *The minimal ruin probability $V(s)$ is a classical solution to the dynamic equation*

$$0 = \inf_A \{\lambda E[V(s-X) - V(s)] + (c + \mu A)V'(s) + A^2\sigma^2 V''(s)/2\},\ s > 0. \qquad (21)$$

The function $V(s)$ has a continuous second derivative $V''(s) < 0$ in $s > 0$, with $\lim_{s\to 0} V''(s) = -\infty$. The optimal amount $A(s)$ invested at state s is $A(0) = 0$ and $A^*(s) = -\mu V'(s)/(\sigma^2 V''(s)),\ s > 0$.

Two different proofs are given in Hipp and Plum (2000, 2003).

There are only a few parameters and exponential claim sizes for which $A(s)$ or $V(s)$ can be given in explicit form.

Example 2 Let $\mu = \sigma = \lambda = 1$ and $c = 3/2$. The claim size has an exponential distribution with mean a. Then

$$A(s) = \sqrt{2c/a}\sqrt{1 - e^{-2as}}.$$

Here, $A(s)/s \to \infty$, and this is a typical behavior of the optimal investment strategy. Since unlimited leverage is forbidden for insurers, leverage has to be bounded or completely forbidden by constraints on the strategies. Such constraints can be defined state dependent, allowing a range $\mathcal{A}(s)$ for the choice of the amount invested at surplus s. With these we can deal with the case of no restriction $\mathcal{A}(s) = (-\infty, \infty)$, no leverage $\mathcal{A}(s) = (-\infty, s]$, no short-selling $\mathcal{A}(s) = [0, \infty)$, neither leverage nor short-selling $\mathcal{A}(s) = [0, 1]$, and bounded leverage and bounded short-

selling $\mathcal{A}(s) = [-as, bs]$. The constraints change the control problem substantially; so, e.g. for no leverage constraint there is no optimal investment strategy. An optimal strategy would be to invest the amount $-\infty$ on the market (volatility hunger). Furthermore, constraints can lead to non-smoothness of the value function.

Such cases are investigated in the papers Azcue and Muler (2010), Belkina et al. (2014), Edalati (2013), Edalati and Hipp (2013), and Hipp (2015). While the proofs and arguments in these papers are all different, it is good to have a universal numerical method (see Sect. 8) which works in all these situations.

The corresponding dynamic equation for the value function reads

$$0 = \inf_{A \in \mathcal{A}(s)} \{\lambda E[V(s-X) - V(s)] + (c + \mu A)V'(s) + A^2\sigma^2 V''(s)/2\}, \ s > 0.$$

If $\mathcal{A}(s) = [a(s), b(s)]$ is an interval, then the minimization with respect to A is easy: there are only three possible minima: at $a(s), b(s)$ or at the unconstrained minimizer $A^*(s)$.

The resulting optimal investment strategies vary according to the claim size distribution. For the unconstrained case, we see that $A^*(s)$ is

1. bounded and converging to $1/R$, the adjustment coefficient of the problem, in the small claims case (Hipp and Schmidli, 2004),
2. unbounded increasing in the large claims case as Weibull, Lognormal, Pareto: the larger risk, the higher the amount invested (Schmidli, 2005)
3. asymptotically linear for Pareto claims.
4. very special when claims are constant (see Sect. 8).

Extensions to other models cause little technical problems. Interest rate earned on surplus or paid for loans can be implemented (see Hipp and Plum 2003).

In the case of two (correlated) stocks, a very simple model would be the dynamics

$$dZ_i(t) = a_i Z_i(t)dt + b_i Z_i(t)dW_i(t), \ t \geq 0, i = 1, 2,$$

where ρ is the correlation between $W_1(t)$ and $W_2(t)$. If we first choose the proportion p and $1 - p$ in which we invest the amount A in stock 1 and stock 2, then we obtain the usual dynamic equation with μ and σ^2 depending on p. Taking the minimum over A the dynamic equation remains with the term

$$-\frac{1}{2}\frac{\mu^2 V'(s)^2}{\sigma^2 V''(s)^2},$$

and since $V'(s), V''(s)$ are fixed, we have to maximize μ^2/σ^2 which produces the well-known optimum

$$p = \frac{a_1 b_2^2 - a_2 \rho}{a_1 b_2^2 + a_2 b_1^2 - (a_1 + a_2)\rho}$$

which is constant. So we indeed have investment into just one index with price process $pZ_1(t) + (1-p)Z_2(t)$.

In other market models for the stock price $Z(t)$, the return on investment will depend on the current price $Z(t)$ which must be included as state variable: for the dynamics

$$dZ(t) = (\mu - Z(t))^2 dt + Z(t)dW(t),\ t \geq 0$$

we will do no or only little investment when $Z(t)$ is close to μ.

Optimal investment can also be used to maximize the company value. This leads to a similar dynamic equation in which changes for dividend payment are necessary (see Azcue and Muler 2010). For simultaneous control of investment and reinsurance, also with constraints, see Edalati (2013).

5 Reinsurance Control

Reinsurance is a most important tool for risk management in insurance. We restrict ourselves on reinsurance of single claims, so we disregard stop loss reinsurance which would ask for a time discrete model. In single claims reinsurance we have a risk sharing between first insurer and reinsurer described by some function $g(x)$ satisfying $0 \leq g(x) \leq x$ which denotes the amount paid by the first insurer; the amount of the reinsurer for a claim of size x is $x - g(x)$. Let G be the set of all risk sharings on the market, and assume that there is $g_0 \in G$ with $g_0(x) = x$ (no reinsurance).

Optimal reinsurance will here be considered for minimizing the first insurer's ruin probability. For maximizing the company value, see Azcue and Muler (2005).

Optimal control for reinsurance is done on a market in which for a risk sharing $g(x)$ a price is specified, and this price determines the optimal strategy. If reinsurance is unaffordable on the market, then it will be optimal not to buy reinsurance. On the other hand, if reinsurance is cheap, then it might be optimal to transfer the total risk to the reinsurer and reach a position with zero ruin probability.

For this exposition of reinsurance control we take a Lundberg model for the risk process.

Assume now that a price system $h(g),\ g \in G$, is given which for each risk sharing defines its reinsurance price, with $h(g_0) = 0$. If at time t the reinsurance contract $g_t(x)$ is active, then the risk process of the first insurer is given by

$$S(t) = s + ct - \int_0^t h(g_u)du - \sum_1^{N(t)} g_{T_i}(X_i),$$

where $T_1, T_2, \ldots$ are the time points at which claims occur. The generator for a fixed risk sharing $g \in G$ is

$$\mathcal{G}^g f(s) = \lambda E[f(s - g(X)) - f(s)] + (c - h(g))f'(s). \tag{22}$$

We are minimizing the infinite horizon ruin probability through dynamic reinsurance, which leads us—as in the discrete case—to a dynamic equation for the control problem, the well-known *Hamilton-Jacobi-Bellman* equation:

$$0 = \inf_{g \in G} \{\lambda E[V(s - g(X)) - V(s)] + (c - h(g))V'(s)\}, \ s \geq 0, \tag{23}$$

with the boundary values $V(\infty) = 0$ and $V(s) = 1, \ s < 0$. Rearranging terms, we obtain

$$V'(s) = \sup_{g \in G: h(g) < c} \frac{\lambda E[V(s) - V(s - g(X))]}{c - h(g)} \tag{24}$$

From this equation we come to the recursion

$$V'_{n+1}(s) = \sup_{g \in G: h(g) < c} \frac{\lambda E[V_n(s) - V_n(s - g(X))]}{c - h(g)}, \tag{25}$$

$$V_{n+1}(s) = -\int_s^\infty V'_{n+1}(x)dx, \tag{26}$$

which produces an increasing sequence of continuous function converging to a solution of (23) when we start with $V_1(s) = \psi(s)$, the infinite time ruin probability without reinsurance. This recursion is, however, not adequate for numerical computations.

In order to obtain a nontrivial solution for our control problem, total reinsurance $g_0(x) = 0$ should be expensive in the sense that $h(g_0) > c$. Otherwise total insurance would be affordable and yield a ruin probability zero for the first insurer.

In this paper, we will restrict ourselves to reinsurance prices computed as a loaded expectation:

$$h(g) = \lambda \rho E[X - g(X)],$$

where $\lambda \rho E[X] > c$.

Common reinsurance forms are

1. proportional reinsurance with $g(x) = bx, \ 0 \leq b \leq 1$,
2. unlimited XL reinsurance with $g(x) = (x - M)^+, 0 \leq M \leq \infty$, and
3. limited XL reinsurance with $g(x) = \min((x - M)^+, L), 0 \leq M, L \leq \infty$.

XL is the usual shorthand for excess of loss. The numbers M and L are called priority and limit, respectively.

Under the above pricing formula, static proportional reinsurance (which is constant over time) does not decrease the first insurer's ruin probability. However in dynamic control, expensive proportional reinsurance can reduce ruin probability.

The unlimited XL reinsurance is optimal for the static situation in the following sense: if $g(x)$ is an arbitrary risk sharing function and $g_M(x)$ an unlimited XL risk sharing with $E[g(X)] = E[g_M(X)]$, then the first insurer's ruin probability with g_M is smaller than the ruin probability with g. Unlimited XL reinsurance is illiquid and/or expensive on reinsurance markets, more common are limited XL forms. Also these have some optimality in the static situation: if g is an arbitrary risk sharing with $x - g(x) \leq L$ and, for some M, $g_{M,L}$ a limited XL reinsurance with $E[g(X)] = E[g_{M,L}(X)]$, then the first insurer's ruin probability with $g_{M,L}$ is smaller than the ruin probability with g.

Optimal dynamic reinsurance strategies take often the position *no reinsurance* when the surplus is small. For proportional reinsurance this was shown by Schmidli (see Schmidli 2007, Lemma 2.14) under the assumption that the price function $h(b)$ satisfies $\liminf_{b\to 0}(c - h(b))/b > 0$.

A similar results holds for unlimited XL reinsurance: If $h(M)$ is continuous at $M = 0$, then there exists $M_0 > 0$ for which $h(M) > c$ for all $0 \leq M \leq M_0$. Choose $s \leq M_0$. The supremum in (24) is taken over $M > M_0 > s$. For $s < M$

$$E[V(s - \min(X, M))] = \mathbb{P}\{X \geq s\} + E[V(s - X)1_{(X\leq s)}]$$

does not depend on M, so the supremum in (24) is attained at $M = \infty$ (no reinsurance) For more details, see Hipp and Vogt (2003).

For limited XL reinsurance, for small surplus s we will see an optimal reinsurance strategy with M and L as well as a price $h(M, L)$ close to but not at zero.

Example 3 We consider a delayed compound Poisson process which has an exponential first waiting time T with mean $\beta = 1$ in which no claims occur, and after time T the claims arrival is a Poisson process with constant intensity $\lambda = 1$. Also the claim sizes have an exponential distribution with a mean 1; the premium rate is $c = 2$. What is the optimal dynamic unlimited XL reinsurance which minimizes the ruin probability?

Volcanos show long waiting times between periods with frequent seismic waves. One could model claims caused by these waves as above.

The standard approach for a solution would be to solve the corresponding Hamilton-Jacobi-Bellman equation (16) for the value function $V(s, p)$, where $p(t)$ is the conditional probability of $\lambda(t) = \lambda$, given $S(u)$, $u \leq t$. But we cannot solve the equation with $V(s, 1)$ as boundary condition since the factor of $V_p(s, p)$ is zero when $p = 1$. Since we know $\lambda(t) = \lambda$ after the first claim, we only need the optimal reinsurance strategy until the first claim. Given no claim up to time t, the function $p(t)$ has derivative given in (14) which yields $p(t) = t/(1 + t)$. We use a nonstationary approach.

This seems to work well for ruin without reinsurance: let $\psi(s)$ be the ruin probability for the uncontrolled Lundberg process with intensity λ, $\psi(s) = \exp(-s/2)/2$. From $E\psi(s - X)] = 2\psi(s)$ and $\psi'(s) = -\psi(s)/2$ we can see that the separation of variables works: for $V(s, t) = f(t)\psi(s)$ the dynamic equation

$$0 = p(t)E[\psi(s - X) - V(s, t)] + cV_s(s, t) + V_t(s, t)$$

yields

$$p(t)(2-f(t))-f(t)+f'(t)=0,\ t\geq 0,\ f(\infty)=1,$$

with the solution

$$f(t)=\frac{1}{2}\frac{1+2t}{1+t}$$

and the value $f(0)=1/2$. So, $V(s,t)=(1+p(t))e^{-s/2}/4$. These exact values can be reproduced numerically with $T=300,\ ds=0.01$ and $dt=0.001$.

With reinsurance we consider the value function $V(s,t)$ and its dynamic equation

$$0=\sup_M\{p(t)E[V_1(s-g_M(X))-V(s,t)]+(2-h(M))V_s(s,t)+V_t(s,t)\},$$

where $g_M(X)=\min(X,M)$ and $h(M)=2\rho E[X-g_M(X)]$ is the reinsurance price for priority M, and $V_1(s)$ is the value function for the problem with constant intensity 1. The optimal priority $M(s,t)$ is derived from maximizing

$$p(t)E[V_1(s-g_M(X))]+(2-h(M))V_s(s,t).$$

For large T we start with $V(s,T)=V_1(s)$, and calculate backwards to $t=0$ using the recursion

$$V(s,t-dt)=V(s,t)+dt\{p(t)E[V(s,t)-V_1(s-g_M(X))]-(c-h(M))V_s(s,t) \qquad (27)$$

in which $M=M(s,t)$ is the optimal priority. The parameter for reinsurance is $\rho=1.1$. Of course, no reinsurance is optimal for all $s\geq 0$ when $t=0$. We see six priority curves $M(s,t), 0.2\leq s\leq 2$, for $t=0.05, 0.025, 0.045, 0.095, 0.17, 300$ (from the right) (Fig. 1). The curves do not intersect; for smaller t we transfer less risk to the reinsurer. In particular, the interval without reinsurance decreases with t from $[0, 1.47]$ to $[0, 0.23]$.

For more general Markov switching models one could perhaps adopt the above approach. Starting with a given initial probability vector at time 0, we can compute the filter $p(t)$ for the time without claim. Assume the vectors $p(t)$ converge to p. Since the control problem with initial distribution p can be solved easily, we can use the corresponding value function $V_0(s)$ as $V(s,\infty)$, so again we would start at some large T instead of ∞, and would compute backward to $t=0$ with the appropriate dynamic equation.

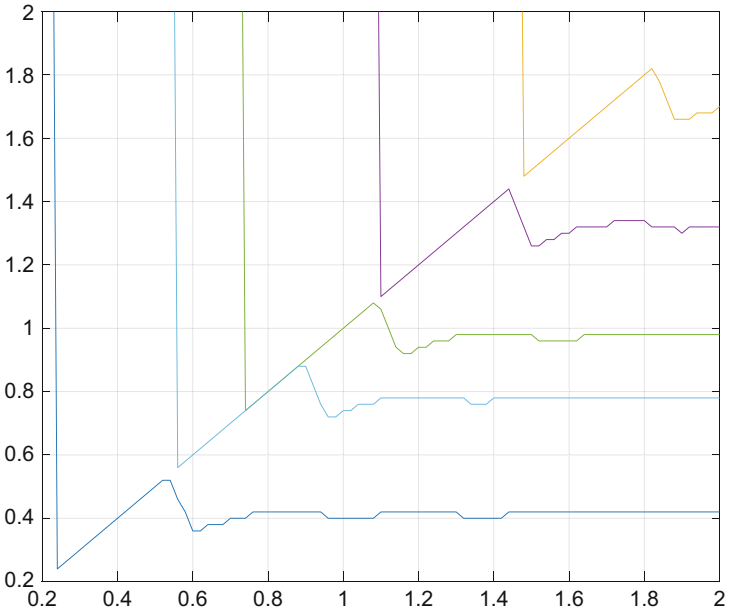


Fig. 1 $M(s,t)$ for $t = 0.005, 0.025, 0.045, 0.095, 0.17, 3005$

6 Dividend Control

Management decisions in insurance, such as reinsurance or investment, have an impact on the company value, and control of investment and reinsurance can be done with the objective to maximize this value. Since the company value is itself the result of a control problem, the maximizing by investment or reinsurance is a control problem with two (or more) control variables, dividends and investment and/or reinsurance. For simplification we restrict ourselves to dividend strategies which are barrier strategies.

Azcue and Muler (in Azcue and Muler 2005, 2010) solve the problems for reinsurance and investment. They mainly characterize the value function as a solution to the dynamic equation, without showing numerical results. For applications in insurance it might be interesting to see whether reinsurance can increase the company value. For reinsurance one has to pay reinsurance premia, and this will reduce the value. But the reduction can be compensated by the reduction of the ruin probability or by increasing the time to ruin for the company. The answer to this question will depend on the relation between premium rate c and reinsurance premia, as well as on the discount rate δ (a large δ reduces the effect of a longer time to ruin). We will present some numerical examples in Sect. 8.

For company values $V(s)$ in a simple diffusion the initial value is $V(0) = 0$. For Lundberg models $V(0)$ is positive and known only in the trivial case when all surplus and premia are paid out, i.e. $V(0) = c/(\lambda + \delta)$ (see Schmidli 2007, Sect. 2.4.2). The

starting value in the general case—with control—can be found exactly as in the case without control: first compute a solution of the dynamic equation $v(s)$ with $v(0) = 1$, and then define the barrier M as $M = \arg\min v'(s)$, and finally

$$V(0) = v(s)/v'(M).$$

For the computation of company values with ruin constraint we suggest the Lagrange multiplier method and a nonstationary approach. For the nonstationary approach we consider dividend payments and ruin probabilities after time t:

$$\begin{aligned} V^D(s,t) &= E\left[\int_t^\infty e^{-\delta u} dD(u) | S(t) = s\right] \\ &\quad -L\mathbb{P}\{S^D(u) < 0 \text{ for some } u \geq t | S(t) = s\}, \\ V(s,t) &= \sup_D V^D(s,t), \\ V(s,\infty) &= -L\psi(s). \end{aligned}$$

Here, $\psi(s)$ is the ruin probability without dividends, and $S^D(u)$ the risk process with dividends which, from time t until time u, add up to $D(u)$. The last relation inspires the following method for computation: start at a large number T, take as initial value the function $V(s,T) = -L\psi(s)$, and then compute backward until $t = 0$ using the non-stationary dynamic equations, modified for dividend payment.

The equations for the backward computation are

$$\begin{aligned} M(t) &= \min\{s : V_s(s,t) = \exp(-\delta t)\}, \\ V(s,t) &= V(s,t+dt) - dt\mathcal{G}V(s,t+dt),\ s \leq M(t), \\ V(s,t) &= V(M(t),t) + (s - M(t))\exp(-\delta t),\ s > M(t). \end{aligned}$$

For a generator involving $V''(s,t)$ which is the case for the simple diffusion model we add $V(0,s) = -L$. For other models we get $V(0,t)$ from $V(0,t+dt)$.

The nonstationary approach deals with partial differential equations for which we most often have to use different discretisations for time and state. The right choice of discretisations is a major problem in the context of these dividend problems (see Sect. 8).

In Sect. 2 an improvement approach was mentioned for the optimal dividend problem with ruin constraint. This was presented in Hipp (2016); however, the method is not sufficiently convincing to be a standard for the numerical computation of the value function in this problem. It might help to find reasonable sub-solutions; it is a method for patient owners of fast computers.

Improvement Approach Assume we have a function $V_n(s,\alpha)$ which is the dividend value for initial surplus s of a strategy which has a ruin probability not exceeding α. We fix $B > s$ and wait without paying dividends until we reach B.

We will reach B before ruin with the probability $A = (1 - \psi(s))/(1 - \psi(B))$, where $\psi(x)$ is the ruin probability without dividends with initial surplus x. At B (we have no upward jumps) we start paying dividends with a strategy corresponding to a ruin probability $a(B)$ having dividend value $V_n(B, a(B))$. The ruin probability of this strategy is $1 - A + Aa(B)$, and the dividend value is the number $V_n(B, a(B))$, discounted to zero. Let τ be the waiting time to reach B, and $v(s)$ be the unique solution of the equation $0 = \delta v(s) + \mathcal{G}v(s)$ with $v(0) = V'(0) = 1$, where δ is the discount rate and $\mathcal{G}$ the generator of the underlying stationary Markov process:

$$0 = \delta v(s) + \lambda E[v(s - X) - v(s)] + cv'(s) \text{ for the Lundberg process}$$
$$0 = \delta v(s) + \mu v'(s) + \sigma^2 v''(s) \text{ for the simple diffusion model.}$$

Then

$$E[\exp(-\delta\tau)] = v(s)/v(B). \tag{28}$$

If we define $a(B)$ from the equation

$$A + (1 - A)a(B) = \alpha,$$

then our dividend strategy has ruin probability α and dividend value

$$V_n(s)v(s)/v(B).$$

For $B \to s$ we obtain the limit $V_n(s, \alpha)$, so a new value dividend function which is an improvement over $V_n(s, \alpha)$ can be defined:

$$V_{n+1}(s, \alpha) = \sup_{B>s} V_n(s, a(B))v(s)/v(B). \tag{29}$$

In each iteration step, we have to compute the V-function for all $s \geq 0$ and $\psi(s) \leq \alpha \leq 1$. And it has to be done on a fine grid. This causes long computation times.

One can start with the function $V_1(s, \alpha) = s - s(\alpha)$, where $s(\alpha)$ is defined through

$$\psi(s(\alpha)) = \alpha.$$

The strategy for this value is: pay out the lump sum $s - s(\alpha)$ at time 0 and stop paying dividends forever. One should also try other initial functions which are closer to the true function, such as $V_1(s, \alpha) = V_0(s - s(\alpha))$ in the simple diffusion model. For the Lundberg model, one can similarly use the function $V_0(s)$, the company value without ruin constraint, but $s(\alpha)$ has to be replaced by a number $s_1(\alpha)$ defined via the equation

$$E[\psi(s - Y)] = \alpha,$$

where Y is the deficit at ruin in the without dividend process. For exponential claims, we can replace Y by X (see Hipp 2016). Notice that s can be smaller than $s_1(\alpha)$, for which the initial value could be $V_1(s) = 0$ or $V_1(s) = s - s(\alpha)$.

7 Viscosity Solutions

In many control problems, the value function can be characterized as the unique viscosity solution to the classical Hamilton-Jacobi-Bellman equation. What is more important: it helps in the proof for convergence of numerical methods (discretizations).

The concept of viscosity solutions—introduced in 1980—is well known today, but still not well enough understood. It is not a subject in most lectures on stochastic processes and control. There are various attempts to make the concept more popular: the famous *User's guide* of Crandall et al. (1992) as well as the books by Fleming and Soner (2006) and Pham (2009). We aim at a better understanding for the concept and properties of viscosity solutions, and its use for the proof of convergence for Euler type discretization schemes of a Hamilton-Jacobi-Bellman equation. This use is based on the fact that upper and lower limits of discretization schemes are viscosity solutions.

In particular we try to provide

1. a better understanding of the Crandall-Ishii maximum principle
2. a proof for the comparison argument which uses $V(0)$ and $V'(0)$
3. an understanding that the concept, being rather technical, is of major importance for applications (numerics and understanding control problems).

For this, we think that a complete and detailed proof for the Crandall-Ishii comparison argument should be included in this section, although for smooth reading one would transfer the proof to an appendix.

Value functions are not always smooth, the viscosity concept is useful to deal with these value functions. Here are two figures from optimization problems with singular value functions; they come from the optimal investment problem with constraint sets $\mathcal{A}(s)$: the amount $A(s)$ invested in stock must lie in $\mathcal{A}(s)$ when we are in state s. In both figures the blue line shows the proportion $A(s)/s$ invested, while the black is the first derivative of the value function $V(s)$ (Figs. 2 and 3).

The dynamic equation for our control problem, valid for $s > 0$, is

$$0 = \sup_{A \in \mathcal{A}(s)} \left\{ \lambda E[V(s-U) - V(s)] + (c + A)V'(s) + A^2 V''(s)/2 \right\}.$$

Because of the above examples there is no hope for the statement: the value function is the unique smooth solution of the above dynamic equation. Instead one can try to prove that the value function is a (unique smooth) viscosity solution of the above HJB. For this section we will always consider the optimal investment problem with

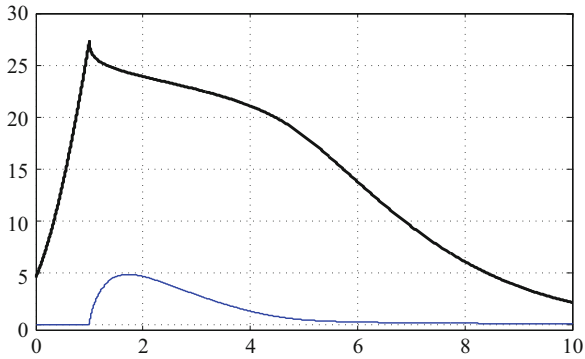


Fig. 2 $\mathcal{A}(s) = \{0\}$, $s < 1, \mathcal{A}(s) = [0, \infty), s \geq 1$

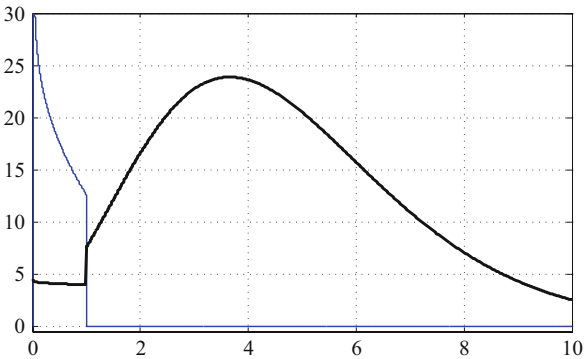


Fig. 3 $\mathcal{A}(s) = [0, \infty)$, $s < 1, \mathcal{A}(s) = \{0\}, s \geq 1$

constraints; in particular, sub-solutions, super-solutions, and viscosity solutions are always defined with respect to the above HJB.

Definition 4 A function $V(s),\ s \geq 0$, is a viscosity **super-solution** at $s > 0$ if for $V(x) \geq \phi(x) \in C_2$ having in s a local minimum for $V(x) - \phi(x)$

$$\sup_{A \in \mathcal{A}(s)} \left\{\lambda E[V(s - U) - V(s)] + (c + A)\phi'(s) + A^2\phi''(s)/2\right\} \leq 0.$$

$V(s)$ is a viscosity **sub-solution** at $s > 0$ if for $V(x) \leq \phi(x) \in C_2$ having in s a local maximum for $V(x) - \phi(x)$

$$\sup_{A \in \mathcal{A}(s)} \left\{\lambda E[V(s - U) - V(s)] + (c + A)\phi'(s) + A^2\phi''(s)/2\right\} \geq 0.$$

$V(s)$ is a **viscosity solution:** if it is a super- and sub-solution at all $s > 0$.
An equivalent definition using sub- and superjets is

Definition 5 $V(s)$ is a viscosity super-solution at $s > 0$ if for $V(s + h) \leq V(s) + ah + bh^2/2 + o(h^2)$ we have

$$\sup_{A\in\mathcal{A}(s)}\left\{\lambda E[V(s-U)-V(s)]+(c+A)a+A^2b/2\right\}\le 0.$$

$V(s)$ is a viscosity sub-solution at $s>0$ if for $V(s+h)\ge v(s)+ah+bh^2/2+o(h^2)$

$$\sup_{A\in\mathcal{A}(s)}\left\{\lambda E[V(s-U)-V(s)]+(c+A)a+A^2b/2\right\}\ge 0.$$

(a,b) are called 2nd order sub- and super-jet of $V(x)$ at s.

The concept of viscosity solutions is important for numerical methods which are based on Euler type discretisations of the dynamic equation. The discretized version of the value function $V_\Delta(s)$ is the numerical solution for step size $\Delta>0$ which, at $s=k\Delta$, is defined from

$$0=\sup_{A\in\mathcal{A}(s)}\left\{\lambda(g_\Delta(s)-V_\Delta(s))+(c+A)V'_\Delta(s)+A^2V''_\Delta(s)/2\right\},$$

with

$$g_\Delta(s)=\sum_{i=1}^{k}V_\Delta((k-i)\Delta)\mathbb{P}\{(i-1)\Delta\le X<i\Delta\}.$$

$$V'_\Delta(s)=(V_\Delta(s+\Delta)-V_\Delta(s))/\Delta,$$

$$V''_\Delta(s)=(V'_\Delta(s)-V'_\Delta(s-\Delta))/\Delta.$$

Its computation is possible via the recursion:

$$V'_\Delta(s)=\inf_{A\in\mathcal{A}(s)}\frac{\lambda\Delta(V_\Delta(s)-g_\Delta(s))+A^2V'_\Delta(s-\Delta)/2}{\Delta(c+A)+A^2/2}$$

Then

$$V^*(x)=\limsup_{s=k\Delta\to x,\Delta\to 0}V_\Delta(s)$$

is a viscosity sub-solution, while

$$V_*(x)=\liminf_{s=k\Delta\to x,\Delta\to 0}V_\Delta(s)$$

is a viscosity super-solution of the dynamic equation. The convergence is a strong convergence concept: it implies uniform convergence on compact sets.

A convergence proof (see Chap. IX of Fleming and Soner 2006) can now be very short: since a sub-solution can never be larger than a super-solution, we have $V^*(x)\le V_*(x)$. Since $V_*(x)\le V^*(x)$, by definition, we have equality. For the above

inequality between the sub- and the super-solution one uses the famous Crandell-Ishii maximum principle which we discuss later. First we give a proof for the sub-solution property for $\limsup\ V^*(s)$:

Proof Let $\phi(x) \in C_2$ for which $V^*(x) - \phi(x)$ has a strict local minimum at s_0. With $\phi_\Delta(s)$ being the restriction of $\phi(x)$ to the Δ-grid we define

$$s_\Delta = \arg\min_{s=k\Delta\geq 0} V_\Delta(s) - \phi_\Delta(s).$$

Then

$$V_\Delta(s_\Delta) - \phi_\Delta(s_\Delta) \leq V_\Delta(s_\Delta \pm \Delta) - \phi_\Delta(s_\Delta \pm \Delta),$$

and so $V'_\Delta(s_\Delta) \leq \phi'_\Delta(s_\Delta)$ and $V''_\Delta(s_\Delta) \leq \phi''_\Delta(s_\Delta)$. We can find a sequence Δ_n for which $s_{\Delta_n} \to s$ and $V_{\Delta_n}(s_{\Delta_n}) \to V^*(s)$. Recall that $V_\Delta(s)$ is a solution to the discretised dynamic equation

$$0 = \sup_{A\in\mathcal{A}(s)} \left\{\lambda(g_\Delta(s) - V_\Delta(s)) + (c + A)V'_\Delta(s) + A^2 V''_\Delta(s)/2\right\}.$$

Now let $s = s_\Delta$ and $\Delta = \Delta_n$ and $n \to \infty$. Then the first term in the brackets has only limits $\leq E[V^*(s-U)]$ (by Fatou's lemma), the second term in the brackets has limit $= -V^*(s)$, the third term in the brackets has limits $\leq (c + A)\phi'(s)$, and the last term in the brackets has limits $\leq A^2/2\ \phi''(s)$. So

$$0 \leq \sup_{A\in\mathcal{A}(s)} \left\{\lambda E[V^*(s-U) - V^*(s)] + (c + A)\phi'(s) + A^2\phi''(s)/2\right\},$$

which is the desired result for a sub-solution. □

The inequality *sub-solution* $\leq$ *super-solution* is based on the famous maximum principle.

Theorem 2 *Assume that* $\mathbb{P}\{U > x\} > 0$ *for all* $x > 0$, *and that the constraints* $\mathcal{A}(x)$ *are intervals* $[a(x), b(x)]$ *with Lipschitz functions* $a(x), b(x)$ *satisfying* $b(x) > 0, x > 0$. *Let* $v(x), w(x)$ *with* $v(0) \leq w(0)$ *be locally Lipschitz,* $v(x)$ *a sub-solution and* $w(x)$ *a super-solution of our dynamic equation. Assume that* $v(x) - w(x)$ *has a strict local maximum in* $(0, \infty)$. *Then* $v(x) \leq w(x)$ *for all* $x \geq 0$.

This statement is concerned with the values $v(x), w(x)$ for $x > 0$. We define $v(x) = w(x) = 0$ for $x < 0$ and note that $\mathbb{P}\{U \leq 0\} = 0$. We shall first give a simple proof for the case that the function $v(x)$ and $w(x)$ have continuous second derivatives.

Proof Simple version: Assume that $v(x), w(x)$ are twice differentiable on $(0, \infty)$ having a global maximum x^* for $v(x) - w(x)$ in $(0, K)$. For $\xi > 0$ let

$$(x_\xi, y_\xi) = \arg\max_{0<x,y<K} v(x) - w(y) - \xi(x - y)^2.$$

Then $v'(x_\xi) = w'(y_\xi) = 2\xi(x_\xi - y_\xi)$, and $v''(x_\xi) = w''(y_\xi) = 2\xi$. For $\xi \to \infty$ we have $x_\xi \to x^*$ and $y_\xi \to x^*$ and furthermore $\xi(x_\xi - y_\xi)^2 \to 0$. Define

$$H_1(A) = E[v(x_\xi - U) - v(x_\xi)] + (c + A)v'(x_\xi) + A^2 v''(x_\xi),$$

$$H_2(A) = E[w(y_\xi - U) - w(y_\xi)] + (c + A)w'(y_\xi) + A^2 w''(y_\xi).$$

Then

$$\sup_{A \in \mathcal{A}(x_\xi)} H_1(A) \geq 0. \text{ and } \sup_{A \in \mathcal{A}(y_\xi)} H_2(A) \leq 0.$$

So there is $A_\xi \in \mathcal{A}(x_\xi)$ and $B_\xi \in \mathcal{A}(y_\xi)$ with

$$|A_\xi - B_\xi| \leq L|x_\xi - y_\xi|$$

where L is the Lipschitz constant, giving

$$\begin{aligned} H_1(A_\xi) - H_2(B_\xi) &= I(1) + I(2) + I(3) \geq 0, \\ I(1) &= \lambda E[v(x_\xi - U) - w(y_\xi - U)] - (v(x_\xi) - w(y_\xi)), \\ I(2) &= (c + A_\xi)v'(x_\xi) - (c + B(y_\xi))w'(y_\xi)), \\ I(3) &= A_\xi^2 v''(x_\xi)/2 - B_\xi^2 w''(y_\xi)/2. \end{aligned}$$

Now

$$\begin{aligned} I(2) &= (A_\xi - B_\xi)2\xi(x_\xi - y_\xi) \leq 2L(x_\xi - y_\xi)^2 \to 0, \ \xi \to \infty. \\ I(3) &\leq 2\xi(A_\xi - B_\xi)^2 \leq 2L^2\xi(x_\xi - y_\xi)^2 \to 0. \end{aligned}$$

With $f(h) = v(x_\xi + Ah) - w(y_\xi + Bh) - \xi(x_\xi + Ah - y_\xi - Bh)^2$ we have $f(h) \leq f(0)$ and so $f''(0) \leq 0$, i.e.

$$A^2 v''(x_\xi) - B^2 w''(y_\xi) \leq 2\xi(A - B)^2.$$

This yields

$$I(1) \to \lambda E[v(x^* - U) - w(x^* - U)] - (v(x^*) - w(x^*))$$

$$\leq M(\mathbb{P}\{U \leq x^*\} - 1) < 0,$$

with $M = v(x^*) - w(x^*)$, a contradiction. □

Here is a proof without derivatives. It is clearly inspired by the proof given in the *User's guide*, with some modifications.

Proof First we restrict the argument x to a finite interval $(0, K)$ containing a global maximum x^* with $v(x^*) - w(x^*) = M > 0$. For $n > 0$ and $0 < x < K$ define

$$v_n(x) = \sup_{\hat{x}\in[0,K]} v(\hat{x}) - n^2(x - \hat{x})^2. \tag{30}$$

These functions are semiconvex (i.e., $v_n(x) + Sx^2$ is convex for some $S > 0$). Similarly, for $n > 0$ we define

$$w_n(y) = \inf_{\hat{y}\in[0,K]} w(\hat{y}) + n^2(y - \hat{y})^2,$$

which is semiconcave ($w_n(y) - Sy^2$ concave for some S). We have

$$0 \le v(x) - v_n(x) \le L^2/n^2 \text{ and } 0 \le w_n(y) - w(y) \le L^2/n^2.$$

The functions $v_n(x), w_n(y)$ are twice differentiable almost everywhere (according to Alexandrov's theorem, see Crandall et al. 1992, Theorem A.2, p. 56, with a 1.5 pp proof).

Now let $\overline{x}, \overline{y}$ be given at which we have second derivatives for $v_n(x), w_n(y)$. Let $\hat{x}$ be the maximizer in (30), i.e. satisfying $v_n(\overline{x}) = v(\hat{x}) - n^2(\overline{x} - \hat{x})^2$, and denote the similar point for $w_n(x)$ and $\overline{y}$ by $\hat{y}$. For notational convenience we omitted the dependence on n.

Then for small enough h we have $v_n(\overline{x} + h) \ge v(\hat{x} + h) - n^2(\overline{x} - \hat{x})^2$ and then

$$v(\hat{x} + h) \le v(\hat{x}) + hv_n'(\overline{x}) + h^2 v_n''(\overline{x})/2 + o(h^2).$$

Similarly,

$$w(\hat{y} + h) \ge w(\hat{y}) + hw_n'(\overline{y}) + h^2 w_n''(\overline{y})/2 + o(h^2),$$

which implies the two inequalities

$$\sup_{A\in\mathcal{A}(\hat{x})} \left\{\lambda E[v(\hat{x} - U) - v(\hat{x})] + (c + A)v_n'(\overline{x}) + A^2 v_n''(\overline{x})/2\right\} \ge 0,$$

$$\sup_{A\in\mathcal{A}(\hat{y})} \left\{\lambda E[w(\hat{y} - U) - w(\hat{y})] + (c + A)w_n'(\overline{y}) + A^2 w_n''(\overline{y})/2\right\} \le 0,$$

Finally we apply Jensen's Lemma for semiconvex functions (Lemma A.3 in Crandall et al. 1992), which in our special situation reads

Lemma 7 *Let $r > 0$ and $\delta > 0$ be arbitrary. Then the set of (x^*, y^*) with $||(x^*, y^*) - (x_\xi, y_\xi)|| < \delta$ for which*

$$v_n(x) - w_n(y) - \xi(x-y)^2 - p_1 x - p_2 y$$

is maximized at (x^*, y^*) *for some* p_1, p_2 *with* $p_1^2 + p_2^2 < r$ *has positive measure.*

For $\xi > 0$ let

$$(x_\xi, y_\xi) = \arg\max_{x,y\in[0,K]}\{v_n(x) - w_n(y) - \xi(x-y)^2\} + p_1 x - p_2 y$$

with $p_1^2 + p_2^2$ small for which the second derivatives of v_n and w_n exist at x_ξ and y_ξ, respectively. (x_ξ, y_ξ) depends on ξ, p, n.

For some $A \in \mathcal{A}(\hat{x}_\xi)$

$$I(A) = \lambda E[v(\hat{x}_\xi - U) - v(\hat{x}_\xi)] + (c + A)v_n'(x_\xi) + A^2 v_n''(x_\xi)/2 \geq 0.$$

For all $B \in \mathcal{A}(\hat{y}_\xi)$

$$I(B) = \lambda E[w(\hat{y}_\xi - U) - v(\hat{y}_\xi)] + (c + B)w_n'(y_\xi) + B^2 w_n''(y_\xi)/2 \leq 0.$$

The difference $I(A) - I(B)$ is non-negative for some $A_\xi \in \mathcal{A}(\hat{x}_\xi)$ and $B_\xi \in \mathcal{A}(\hat{y}_\xi)$ satisfying

$$|A_\xi - B_\xi| \leq L|\hat{x}_\xi - \hat{y}_\xi|.$$

We now let $p \to 0, n \to \infty, \xi \to \infty$ in this order! The difference consists of three terms:

$$\begin{aligned}
I(1) &= E[v(x_\xi - U) - v(x_\xi)] - E[w(y_\xi - U) - w(y_\xi)],\\
I(2) &= (c + A_\xi)v_n'(x_\xi) - (c + B_\xi)w_n'(y_\xi),\\
I(3) &= B_\xi^2 v_n''(x_\xi)/2 - A_\xi^2 w_n''(y_\xi)/2\\
v_n'(x_\xi) &= 2\xi(x_\xi - y_\xi) + p_1,\\
w_n'(y_\xi) &= 2\xi(x_\xi - y_\xi) + p_2.
\end{aligned}$$

We have

$$|I(2)| \leq c||p|| + 2\xi|\hat{x}_\xi - \hat{y}_\xi||x_\xi - y_\xi|$$

converges to zero for $p \to 0, n \to \infty, \xi \to \infty$.

The argument in the proof with second derivatives leads to

$$|I(3)| \leq 2\xi(A_\xi - B_\xi)^2 \leq 2L^2\xi(\hat{x}_\xi - \hat{y}_\xi)^2$$

which converges to 0 for $\xi \to \infty$.

Finally, with $x_\xi \to x^*$ and $y_\xi \to x^*$

$$I(1) \to E[v(x^* - U) - w(x^* - U)] - (v(x^*) - w(x^*)) < 0$$

because of $v(x) - w(x) \le v(x^*) - w(x^*) = M$ and

$$E[v(x^* - U) - w(x^* - U)] \le M\mathbb{P}\{U \le x^*\} < M.$$

This contradicts that the difference must be non-negative, so $M > 0$ cannot be true, and thus our assertion $v(x) \le w(x), x \ge 0$, holds. □

Usually, the maximum principle is applied for $v(0) = w(0)$ and $v(\infty) = w(\infty)$, so the initial conditions are for values of the functions. This is appropriate for diffusion models where we often have $v(0) = w(0) = 0, v(\infty) = w(\infty) = 1$.

In Lundberg models we have instead $v(\infty) = w(\infty) = 1$ and a given value for the derivative at zero:

$$v'(0) = -\lambda(1 - v(0))/c, w'(0) = -\lambda(1 - w(0))/c.$$

Fortunately, with the above maximum principle one can also handle this situation.

Lemma 8 *Assume that* $\mathbb{P}\{U > x\} > 0$ *for all* $x > 0$, *and that the constraints* $\mathcal{A}(x)$ *are intervals* $[a(x), b(x)]$ *with Lipschitz functions* $a(x), b(x)$ *satisfying* $b(x) > 0$, $x > 0$.

Let $v(x), w(x)$ *be viscosity solutions of our dynamic equation having continuous first derivatives with* $v(0) = w(0)$ *and* $v'(0) = w'(0)$. *Then* $v(x) = w(x)$ *for all* $x \ge 0$.

Proof Assume that there exists $x_0 \ge 0$ such that $v(x) = w(x), 0 \le x \le x_0$ and that $v(x) < w(x)$ for $x_0 < x \le x_0 + \varepsilon$. The case $v(\infty) \ge w(\infty)$ is easy.
Assume $v(\infty)(1 + \gamma)^2 < w(\infty)$.
Choose $x_2 > x_0$ close to x_0 such that $v'(x)(1 + \gamma) \ge w'(x), 0 \le x \le x_2$. Define

$$V(x) = w(x), x \le x_2, \text{ and } V'(x) = v'(x)(1 + \gamma), x \ge x_2.$$

Similarly,

$$W(x) = v(x), x \le x_2, \text{ and } W'(x) = w'(x)/(1 + \gamma), x \ge x_2.$$

with the properties

1. $V(0) = W(0)$,
2. $V(x), W(x)$ are Lipschitz,
3. $V(x)$ is a sub-solution and $W(x)$ a super-solution,
4. $V(\infty) \le W(\infty)$.

Hence

$V(x) \le W(x), x \ge 0,$ rm contradicting $V(x_2) = w(x_2) > v(x_2) = W(x_2)$. □

For the above discretization schemes, one can prove equi-continuity of the approximations $V'_\Delta(s), s = k\Delta \ge 0$ (see Hipp 2015) which implies that lim sup and lim inf have continuous derivatives.

In all, we can prove that the discretization schemes converge to some function $W(x)$ having a continuous first derivative. For many optimization problems one can also show that the value function $V(x)$ is a viscosity solution of the corresponding HJB equation. However, we need a continuous first derivative for $W(x)$ to obtain $V(x) = W(x)$ from the above comparison argument. It is still open for which optimization problem the value function $V(x)$ has a continuous derivative. So, regrettably, we do not know whether the limit of our discretizations is the value function of the given control problem.

Cases in which the value function is known to have a continuous first and second derivative are

- unrestricted case: $\mathcal{A}(x) = (-\infty, \infty)$ (see Hipp and Plum 2003),
- no short-selling and limited leverage: $\mathcal{A}(x) = [0, bx]$ (see Azcue and Muler 2010),
- bounded short-selling and bounded leverage: $\mathcal{A}(x) = [-ax, bx]$ (see Belkina et al. 2014)

8 Numerical Issues

Numerical computations for solutions of control problems are demanding, they cannot be done on a simple spreadsheet. The results shown in this article are all done with MatLab. This matrix oriented programming language is well suited for the handling of large arrays; in particular, the commands `find` and `cumsum` (or `cumtrapz`) are used frequently, and arrays with more that a million entries were stored and handled easily.

Continuous time and state functions have to be discretized, and the same is done with integrals and derivatives. The step size for the surplus will be denoted by ds and for time by dt. If other state variables show up in the model (e.g., in mixture models), we try to replace them by t in a nonstationary model. We will use Euler type discretisations of the following kind: with $s = k\,ds$

$$V_s(s,t) = (V(s+ds,t) - V(s,t))/ds,$$
$$V_{ss}(s,t) = (V_s(s,t) - V_s(s-ds))/ds,$$
$$V_t(s,t) = (V(s,t+dt) - V(s,t))/dt,$$

$$E[V(s-X,t)] = \sum_{i=1}^{k} V(s-i\,ds)\mathbb{P}\{(i-1)ds \le X < i\,ds\}.$$

For the expectation, one could use higher order integration methods; however, we here essentially need summation with weights which add up to 1.

In most control problems, the difference between maximizing survival probability and maximizing company value is very small: Rearranging the dynamic equation to solve for $V'(s)$, we obtain in the reinsurance control problem

$$V'(s) = \min \frac{\lambda V(s) - \lambda E[V(s-g_a(X))]}{c - h(a)} \quad \text{for survival prob.}$$

$$V'(s) = \min \frac{(\lambda+\delta) V(s) - \lambda E[V(s-g_a(X))]}{c - h(a)} \quad \text{for dividends.}$$

Since the equations are homogeneous, one can use an arbitrary value for $V(0)$ to see the optimal strategy.

Reinsurance Example In our first example we consider optimal unlimited XL reinsurance for a Lundberg model, first for maximizing the company value, and second to minimize the ruin probability. The parameters are $\lambda = 1$, $c = 2$, $\delta = 0.07$, and the claims have an exponential distribution with mean 1. First we show the derivative of the function $v(s)$ solving the dynamic equation, and next you see the optimal priority $M(s)$ (middle). On the right you see the optimal $M(s)$ which minimizes ruin probability. We see that $v(s)$ has one minimum which is at $M = 4.84$. So the possible values of s are $[0, M]$. In both cases we have a region of small s in which no reinsurance is optimal. Then we see a region with $M(s) = s$, which means reinsurance for the next claim. Then $M(s)$ is increasing almost linearly for the dividend case, while for the ruin case $M(s)$ is almost constant. In both cases, reinsurance is paid for, and in the dividend case this starts at larger surplus. Furthermore, $M(s)$ is higher in the dividend case (which means less reinsurance) (Figs. 4, 5, and 6).

In most optimization problems, the optimizers are found by complete search. In problems with more than one control parameter one should check wether the optimal parameters are continuous in s. Then one can speed up the search: restrict the search for state s on a neighborhood of the value for $s - ds$.

The numerically demanding term is the expectation in the dynamic equation: $E[V_n(s - g_M(X))]$. It has to be computed for many s and M's, and for each iteration n. In some cases this nonlocal term can be transformed to a local one (e.g., for exponential or phase-type distributions,) but with MatLab one can produce the values—following the MatLab-rule `no loops`—in one line. Once define the matrix P of probabilities with step size ds, range $0, ds, 2ds, .., KS\ ds$ for s, and $f(i) = \mathbb{P}\{(i-1)ds \le X < i\,ds\}, i = 1, .., KS$ as

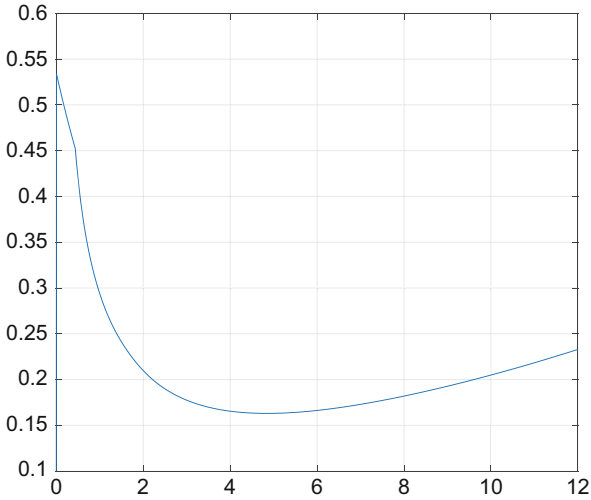


Fig. 4 Derivative of HJB-solution $v'(s)$

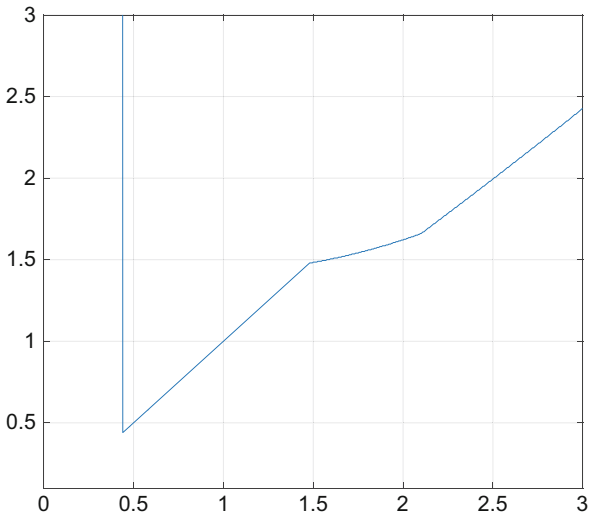


Fig. 5 Optimal priority dividends

$$P(i,j) = f(j), j = 1, \ldots, i-1,$$
$$P(i,i) = \sum_{j=i}^{KS} f(j),$$
$$P(i,j) = 0, j = i+1, \ldots, KS.$$

If $A = \{1 \leq i \leq KS : h(i\,ds) < c\}$, then the vector VI with entries

$$E[V(s-M)] : M = i\,ds,\ i \in A,$$

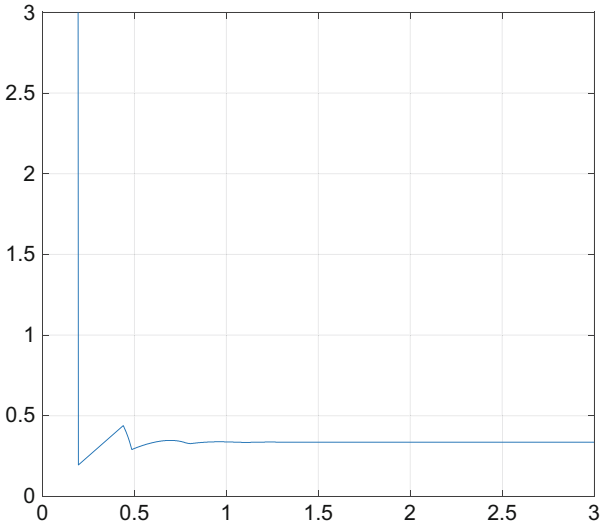


Fig. 6 Optimal priority ruin

is generated by

$$VI = (P(A, 1 : (i-1)) * V(i-1 : -1 : 1)');$$

and the dynamic equation for the value function V in the Lundberg model leads to the formula

$$[V'(s), sb] = \min(\lambda * (V(i) - VI')/(c - h(A)));$$

In special cases the set A can be replaced by a smaller set which speeds up computation.

Investment Example Optimal investment for minimal ruin probability in the Lundberg model has the following equation (where we set $\mu = \sigma^2 = 1$) :

$$0 = \sup_A \lambda E[V(s-X) - V(s)] + (c + A)V'(s) + A^2 V''(s)/2,$$

which has maximizer $A(s) = -V'(s)/V''(s)$. With $U(s) = A^2(s)$ we obtain the equation

$$V'(s) = \frac{\lambda E[V(s) - V(s-X)}{c + \sqrt{U(s)}/2}.$$

For $U(s)$ we get in the case of exponential claims with parameter θ

$$U'(s) = \sqrt{U(s)}(\lambda + 1/2 - \theta c - \theta\sqrt{U(s)}/2) + c$$

(see Hipp and Plum 2003, Remark 8). To obtain the optimal strategy, we can restrict ourselves on $U(s)$ and start with $U(0) = 0$. For the dividend objective we just have to replace λ by $\lambda + \delta$. In the special case $\theta = 1, c = \lambda + 1/2$ we can see that for the dividend objective investment is higher than for the ruin probability objective: for dividends we obtain

$$U'(s) = c - U(s)/2,$$

while for dividends it reads

$$U'(s) = c - U(s)/2 + \delta\sqrt{U(s)}.$$

The above system of two coupled differential equations enables a simple, robust, and efficient computation. The resulting strategies never use short selling, the amount invested $A(s)$ is not always increasing, and generally: the more risky the insurance business is, the larger $A(s)$ will be.

Optimal Investment with Constraints In the unconstrained case, optimal investment is completely different from the one in the unrestricted case. The following figures are based on a Lundberg model with exponential claims for which the unconstrained optimal strategy is increasing and concave and almost constant for large surplus. In the case without leverage and shortselling in the next figure, we see the proportion $A(s)/s$ and the second derivative of the value function. For small s we see $A(s) = s$, and the value function is not concave. An example with volatility hunger is seen in the next figure: here we have the same model and the constraints $\mathcal{A}(s) = [-4s, s]$ (see Belkina et al. 2014). For very small s we have $A(s) = s$, then in a larger interval $A(s) = -4s$, and then the strategy switches back to $A(s) = s$ and continues continuously. The jump from maximal long to a maximal short position can be explained by the fact that a high volatility position can produce also high up movements. The black curve is again the second derivative of the value function.

Constraints can generate singularities in the value function, even the first derivative can have a jump. Such singularities are present also in uncontrolled ruin probabilities, when the claim size distribution has atoms. An example with $X = 1$ is given in the third figure below, it shows $A(s)$ in the unconstrained case (blue line) and for $\mathcal{A}(s) = [0, s]$ (Figs. 7, 8 and 9).

Optimal Dividends with Ruin Constraint The method for the computation of company values with ruin constraint has been described before; we will here discuss the numerical problems and results for the computation using Lagrange multipliers and the nonstationary approach. Our backward calculation starts with $V(s, T) = -L\psi(s)$ which will produce good approximations if T is large enough such that

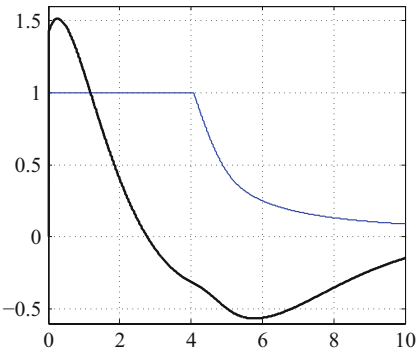


Fig. 7 Constrained optimal investment

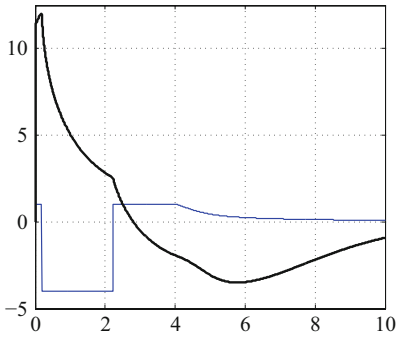


Fig. 8 Example with extreme jumps

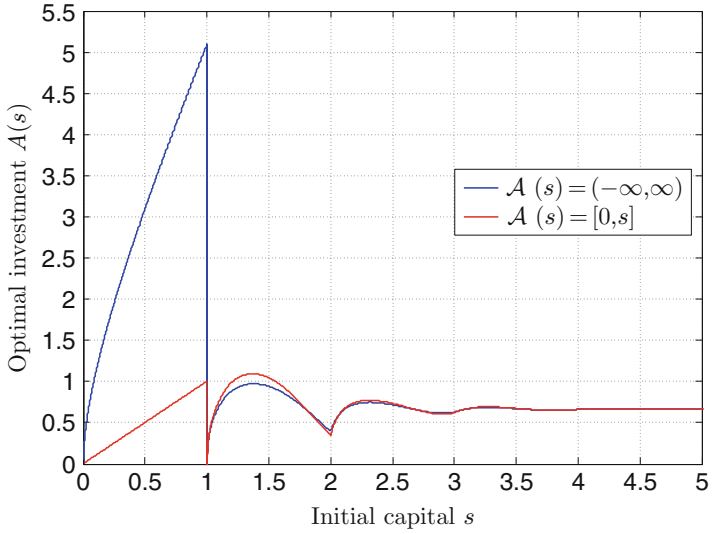


Fig. 9 Optimal investment for $X = 1$

dividend payments do not matter after time T since they are discounted by $e^{-\delta T}$ at least. But the discretization dt must be quite small to get convergence: in the simple diffusion model, for $ds = 0.02$ we need a step size dt of at most 0.0004; for $ds = 0.02$ and $dt = 0.00041$ we obtain results which are completely wrong: barrier close to zero and value functions close to $V(s) = s - L$. The Lundberg model is less sensitive: it works with $ds = 0.02$ and $dt = 0.004$.

The next two figures show the results for simple diffusion models. First, we show the computed curves $V(s, t)$ for 21 values of t, where the largest values belong to $t = 0$. The second is the curve of barriers $M(t)$ which has the expected form: increasing, asymptotically linear, with a decrease close to T. The same form had been obtained in the discrete case of Hipp (2003), the decrease is caused by the choice of $V(s, T)$. The parameters for the plots are $\mu = \sigma^2 = 1$ and the discount rate $\delta = 0.03$.

The third figure shows an efficiency curve for company values and ruin probabilities, which is the same as a plot for $V(s, \alpha)$, the maximal dividend value with a ruin constraint of α. For this we computed $V(s, L)$ with the corresponding ruin probabilities, and plotted the results for a number of L's from 0 to 100. The plot is given for a simple diffusion model with $\sigma = \mu = 1$ and $\delta = 0.07$. The initial surplus is 5. We could not produce reliable results for larger L since they produce $\alpha = 1$ or $\alpha < \psi(5)$. Surprisingly, the dividend value stays near the unconstrained value $V_0(5) = 16.126$ over a long range for α (Figs. 10, 11, and 12).

Results for the Lundberg model are given in the contribution (Hipp, 2016) in these proceedings.

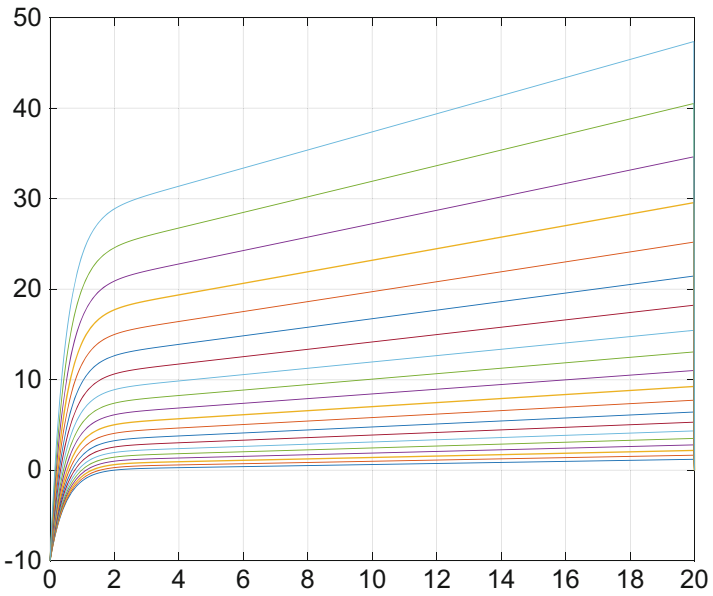


Fig. 10 $V(s, t)$ for $0 \leq s \leq 20$ and various values of t

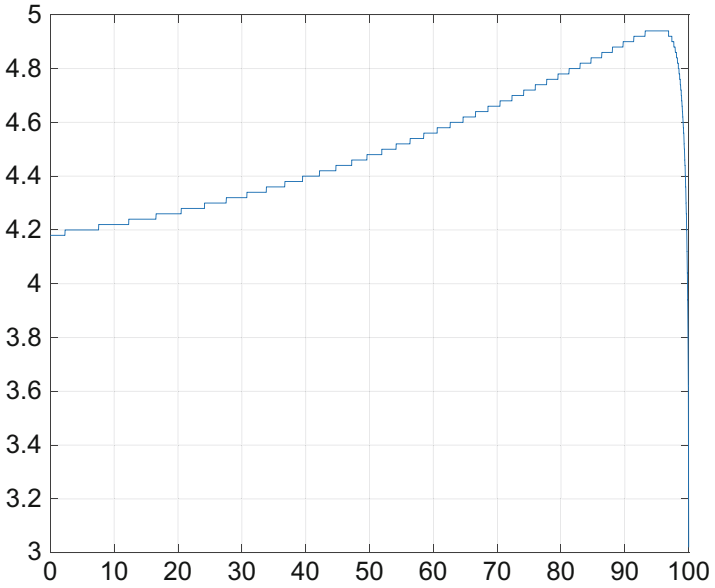


Fig. 11 Optimal barrier $M(t)$

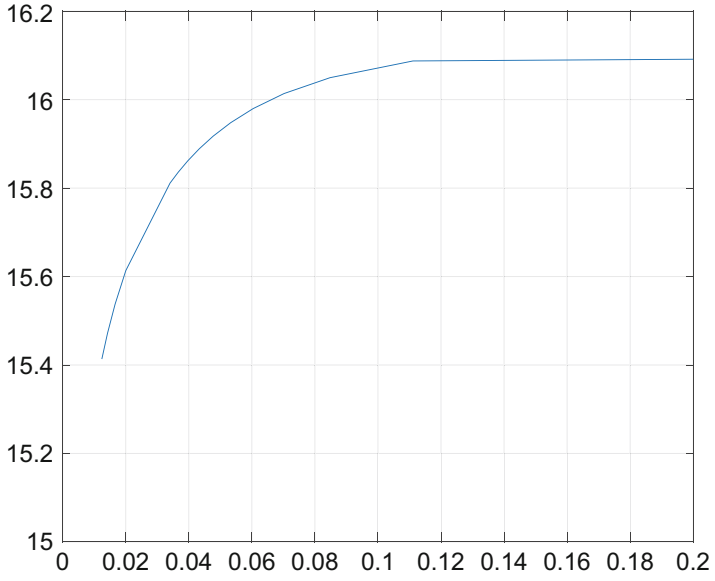


Fig. 12 $V(s,\alpha)$ for $s = 5,\ \psi(s) \leq \alpha \leq 1$

9 Open Problems

Here is a collection of questions which—according to my knowledge—are open, and in my opinion interesting enough to attract (young) mathematicians. They are of course biased by my preferences, but they might still be of some use. They are given in a numbered list, where the order is random (no ordering according to difficulty or importance).

1. For the proof that the discretisations converge to the value function in the optimal investment problem with constraints, one needs that the value function has a continuous derivative. What is the class of problems for which the value function has this property?
2. Optimal company values with ruin constraint are computed with the Lagrange multiplier approach. Do we have a Lagrange gap here? Some positive results are in Hernandez and Junca (2015, 2016).
3. Optimal investment is considered here in a market with constant parameters. How do the solutions change if the market values change as in a finite Markov chain with fixed or random transition rates? What changes if also negative interest is possible?
4. What is the right model for simultaneous control of stop loss and excess of loss reinsurance?
5. Can the nonstationary approach solve control problems also in more complex Markov switching models?
6. Is the capital $V(s, \alpha)$ with ruin constraint a smooth function of s and α?
7. Existing results in models with capital injections solicit the question whether classical reinsurance is still efficient. What is the right model for this question, and what is the answer?
8. Does the approach described at the end of Chap. 5 work for a model in which the state λ is not absorbing?
9. Can the improvement approach in Chap. 6 be applied in the Lundberg model with claim size not exponentially distributed?

References

Albrecher, H., Asmussen, S.: Ruin Probabilities. World Scientific, Singapore (2010)

Azcue, P., Muler, N.: Optimal investment policy and dividend payment strategy in an insurance company. Ann. Appl. Probab. **20**(4), 1253–1302 (2010)

Azcue, P., Muler, N.: Optimal reinsurance and dividend distribution policies in the Cramér-Lundberg model. Math Financ. **15**(2), 261–308 (2005)

Belkina, T., Hipp, C., Luo, S., Taksar, M.: Optimal constrained investment in the Cramer-Lundberg model. Scand. Actuar. J. **2014**(5), 383–404 (2014)

Browne, S.: Optimal investment policies for a firm with a random risk process: exponential utility and minimizing the probability of ruin. Math. Oper. Res. **20**(4), 937–958 (1995)

Crandall, M.G., Ishii, H., Lions, P.L.: User's guide to viscosity solutions of second order partial differential equations. Bull. Am. Math. Soc. **27**(1), 1–67 (1992)

De Finetti, B.: Su un impostazione alternativa della teoria collettiva del rischio. In: Transactions of the XVth International Congress of Actuaries, vol. 2, pp. 433–443 (1957)
Edalati, A.: Optimal constrained investment and reinsurance in Lundberg insurance models. Thesis, KIT (2013)
Edalati, A., Hipp, C.: Solving a Hamilton-Jacobi-Bellman equation with constraints. Stochastics **85**(4), 637–651 (2013)
Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions. Springer, New York (2006)
Hernandez, C., Junca, M.: Optimal dividend payments under a time of ruin constraint: exponential claims. Insur. Math. Econ. **65**, 136–142 (2015)
Hernandez, C., Junca, M.: A time of ruin constrained optimal dividend problem for spectrally one-sided Lévy processes. Submitted (2017). arXiv 1608.02550v2
Hipp, C.: Optimal dividend payment under a ruin constraint: discrete time and state space. Blätter der DGVFM **26**, 255–264 (2003)
Hipp, C.: Correction note to: solving a Hamilton-Jacobi-Bellman equation with constraints. Stochastics **88**(4) 481–490 (2015)
Hipp, C.: Dividend payment with ruin constraint. In: Festschrift Ragnar Norberg. World Scientific, Singapore (2016)
Hipp, C.: Stochastic control for insurance: new problems and methods. In: Proceedings 2nd ICASQF Cartagena (2016)
Hipp, C., Plum, M.: Optimal investment for insurers. Insur. Math. Econ. **27**(2), 215–228 (2000)
Hipp, C., Plum, M.: Optimal investment for investors with state dependent income, and for insurers. Financ. Stoch. **7**(3), 299–321 (2003)
Hipp, C., Schmidli, H.: Asymptotics of the ruin probability for the controlled risk process: the small claims case. Scand. Actuar. J. **2004**(5), 321–335 (2004)
Hipp, C., Vogt, M.: Optimal dynamic XL-reinsurance for a compound Poisson risk process. ASTIN Bull. **33**(2), 193–207 (2003)
Loeffen, R.L.: On optimality of the barrier strategy in de Finettis dividend problem for spectrally negative Lévy processes. Ann. Appl. Probab. **18**(5), 1669–1680 (2008)
Pham, H.: Continuous-Time Stochastic Control and Optimization with Financial Applications. Stochastic Modelling and Application of Mathematics, vol. 61. Springer, New York (2009)
Schmidli, H.: Optimal proportional reinsurance policies in a dynamic setting. Scand. Actuar J. **2001**(1), 55–68 (2001)
Schmidli, H.: On optimal investment and subexponential claims. Insur. Math. Econ. **36**(1), 25–35 (2005)
Schmidli, H.: Stochastic Control in Insurance. Springer, New York (2007)

Stochastic Control for Insurance: New Problems and Methods

Christian Hipp

Abstract Stochastic control for insurance is concerned with problems in insurance models (jump processes) and for insurance applications (constraints from supervision and market). This leads to questions of the following type:

1. How to find numerically a viscosity solution to an integro differential equation;
2. Uniqueness of viscosity solutions when boundary conditions are values of derivatives; and
3. How to solve control problems with the two objectives: dividends and ruin.

We shall present simple Euler schemes (similar to the ones in Fleming–Soner (Controlled Markov Processes and Viscosity Solutions. Stochastic Modelling and Applied Probability. Springer, New York, 2006), Chap. IX) which converge when the value function has a continuous first derivative. This method works in many univariate control problems also when value functions are without continuous second (and first) derivative. Cases with non-smooth value function arise when constraints are restrictive. Furthermore, we consider the infinite horizon problem: maximize dividend payment and minimize ruin probability. This problem is described and solved with a non-stationary approach in the classical Lundberg model.

Keywords Stochastic control • Viscosity solutions • Euler type discretisations • Multi objective problem

AMS Keys Primary 91B30, 93E20; Secondary 49l20, 49L25, 49M25

C. Hipp (Retired)
Karlsruhe Institute of Technology, Institute for Finance and Insurance, Karlsruhe, Germany
e-mail: FChristian.Hipp@gmail.com

J.A. Londoño et al. (eds.), *Actuarial Sciences and Quantitative Finance*, Springer
Proceedings in Mathematics & Statistics 214, https://doi.org/10.1007/978-3-319-66536-8_5

1 Introduction

Stochastic control in finance started more than 40 years ago with Robert Merton's papers *Lifetime portfolio selection under uncertainty: the continuous-time case* (Merton 1969) and *Optimum consumption and portfolio rules in a continuous-time model* (Merton 1971), paving the ground for the famous option pricing articles by Robert Merton *Theory of rational option pricing* (Merton 1973) as well as Fischer Black and Myron Scholes *The pricing of options and corporate liabilities* (Black and Scholes 1973). By now, this is a well-established field with standard textbook such as Fleming-Rishel *Deterministic and Stochastic Optimal Control* (Fleming and Rishel 1975) and Fleming-Soner *Controlled Markov Processes and Viscosity Solutions* (Fleming and Soner 2006), as well as Merton *Continuous Finance* (Merton 1992). I would also mention Karatzas-Shreve *Methods of Mathematical Finance* (Karatzas and Shreve 1998) and the work of Bert Øksendal (2005) and Jerome Stein (2012), and this list is still far from being complete.

Surprisingly, the development of stochastic control in insurance took much longer, although the idea was present already in 1967. Karl Borch (NHH Bergen, Norway) wrote in his *The theory of risk* (Borch 1967, p. 451):

The theory of control processes seems to be tailor made for the problems which actuaries have struggled to formulate for more than a century. It may be interesting and useful to meditate a little how the theory would have developed if actuaries and engineers had realized that they were studying the same problems and joined forces over 50 years ago. A little reflection should teach us that a highly specialized problem may, when given the proper mathematical formulation, be identical to a series of other, seemingly unrelated problems.

As the beginning of stochastic control in insurance one might choose the year 1995 in which Sid Browne's paper *Optimal investment policies for a firm with a random risk process: exponential utility and minimizing the probability of ruin* (Browne 1995) appeared. Since then, this field is very active, and its group of researchers is still growing. A first monograph is Hanspeter Schmidli's book *Stochastic Control in Insurance* (Schmidli 2007) in which an extended list of references contains also earlier work. New books were written recently by Pham (2009) and Azcue and Muler (2014).

Stochastic control in insurance is concerned with control of investment, reinsurance, exposure, and product design. An objective is often the ruin probability which is a dynamic risk measure used in internal models. Minimizing ruin probability results in the reduction of solvency capital, so optimal strategies have also an economic impact. These strategies can be used in scenario generators for management decisions.

Ruin probabilities are not satisfactory when they lead to the decision to stop insurance business (which might happen in reinsurance control with interest on the reserves, or in the control of exposure). Alternatively, one can maximize the value of the company, which is the expected sum of discounted dividends. This objective is more complex, since the company value is itself the result of a control problem: one uses optimal dividend payment.

Company values have the drawback of certain ruin: if an insurer pays dividends to maximize the company value, then the with dividend ruin probability equals 1, no matter how large the initial surplus is. As an alternative we investigate a company value which has a constrained ruin probability. In this setup, only those dividend payments are allowed which lead to a given with dividend ruin probability. This quantity is even more complex since its computation involves a control problem with two objectives, its solution is work in progress.

In this paper we first consider control of constraint investment (such as no short-selling and/or no leverage) to minimize the ruin probability and present the numerical methods for the value function and optimal strategy. Next, concepts and numerical methods are presented for the computation of a company value which has a constrained ruin probability.

2 Optimal Investment for Insurers

Investment of a fixed proportion of the surplus leads to a substantial increase in ruin probability (see Kalashnikov and Norberg 2002). Optimal investment control with unconstrained investment was first given in Hipp and Plum (2000, 2003) for the classical Lundberg model. The risk process at time t is given by

$$S(t) = s + ct - X_1 - \ldots - X_{N(t)},$$

where s is the initial surplus, c the premium intensity, and $X, X_1, X_2, \ldots$ are independent identically distributed claim sizes which are independent of the claims arrival process $N(t)$ being modeled as a homogeneous Poisson process with claim frequency λ. The dynamics of the asset for investment is logarithmic Brownian motion

$$dZ(t) = \mu Z(t)dt + \sigma Z(t)dW(t), t \geq 0,$$

with a standard Wiener process independent of $S(t), t \geq 0$, and constants $\mu, \sigma > 0$. The classical Hamilton-Jacobi-Bellman equation for the minimal ruin probability $V(s)$ reads

$$0 = \inf_A \{\lambda E[V(s - X) - V(s)] + (c + A\mu)V'(s) + A^2\sigma^2 V''(s)/2\}, \; s \geq 0, \qquad (1)$$

where the infimum is taken over all real A representing the amount invested at surplus s. The minimizer $A(s)$ defines an optimal investment strategy given in feedback form: invest the amount $A(s)$ when surplus is s. When X has a continuous density, Eq. (1) has classical bounded solutions, and the unique solution $V(s)$ with $V(\infty) = 0, V'(0) = \lambda(V(0) - 1)/c$ is the minimal ruin probability. The optimizer $A(s)$ converges to zero for $s \to 0$ at the rate $\sqrt{s}$ which leads to $A(s)/s \to \infty$, and this shows that the corresponding investment strategy has unlimited leverage.

Since unlimited leverage strategies are not admissible for insurers, we have to restrict the set of investment strategies for each surplus $s : A \in \mathcal{A}(s)$. Possible restrictions are $\mathcal{A}(s) = (-\infty, s]$ for no leverage, $\mathcal{A}(s) = [0, \infty)$ for no short-selling, or $\mathcal{A}(s) = [0, s]$ for neither leverage nor short-selling. Such constraints change the nature of the control problem: the constraint $\mathcal{A}(s) = (-\infty, s]$ results in a control problem having no solution; other constraints yield a Hamilton-Jacobi-Bellman equation

$$0 = \sup_{A \in \mathcal{A}(s)} \{\lambda E[V(s-X) - V(s)] + (c + A\mu)V'(s) + A^2\sigma^2 V''(s)/2\}, \; s \geq 0, \qquad (2)$$

which does not have a solution with (continuous) second derivative.

For this situation one can use the concept of viscosity solutions described in Fleming and Soner (2006). This concept is tailor made for risk processes which are diffusions (in which ruin probabilities at zero are 1), or for dividend maximization. For ruin probabilities in Lundberg models it has to be modified: instead of two fixed values ($V(0) = 1, V(\infty) = 0$) we have boundary conditions on $V(\infty)$ and $V'(0)$. But also for this situation, the Crandall-Ishii comparison argument is valid under the additional hypothesis that the viscosity solutions to be compared have a continuous first derivative (see Hipp 2015).

The numerical solution of Eq. (2) is done with Euler type discretisations. For a step size Δ we define the function $V_\Delta(s)$ as the solution of the discretised equation, with discretised derivatives

$$V'_\Delta(s) = (V_\Delta(s) - V_\Delta(s-\Delta))/\Delta,$$

$$V''_\Delta(s) = (V'_\Delta(s+\Delta) - V'_\Delta(s))/\Delta,$$

and also the integral is discretised: for $s = k\Delta$ it is approximated by

$$G_\Delta(s) = \sum_{i=1}^{k} V_\Delta(s - i\Delta)\mathbb{P}\{(i-1)\Delta \leq X < i\Delta\}.$$

Example 1 Let $\mathcal{A}(s) = [-bs, as]$ with $a = 1, b = 60$. The other parameters are $c = 3.5, \lambda = 1, \mu = 0.3, \sigma = 1$, and the claim size has an Erlang(2) distribution with density $x \to xe^{-x}, x > 0$. Here, the supremum in (2) can be either the unrestricted maximum $A(s) = -\mu V'(s)/(\sigma^2 V''(s))$ or one of the two values $-bs, as$, whichever produces the smaller value for $V'(s)$. For a possible maximizer A in (2), $A \in \{-bs, as\}$, we have

$$V'_\Delta(s) = \frac{\lambda(V_\Delta(s) - G_\Delta(s)) - 0.5A^2\sigma^2 V'_\Delta(s-\Delta)}{(c + A\mu)\Delta + 0.5A^2\sigma^2}, \qquad (3)$$

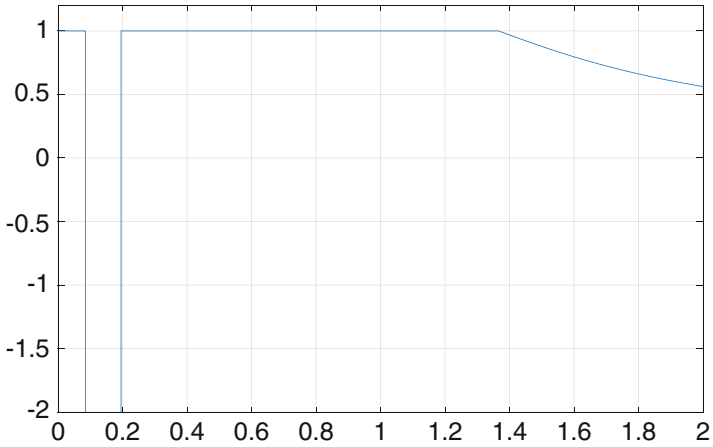


Fig. 1 Optimal proportion $A(s)/s$ invested

and for the unrestricted maximizer we obtain the following quadratic equation with $H(s) = \lambda(G_\Delta(s) - V_\Delta(s))$:

$$V'_\Delta(s)^2(c + 0.5\Delta) = V'_\Delta(s)(H(s) - cV'(s - \Delta)) - H(s)V'(s - \Delta). \tag{4}$$

We see (Fig. 1) that the optimal investment strategy jumps from the maximal admissible long position s to the maximal short position $-60s$, and back. In contrast to the Example 5.1 in Belkina et al. (2012) we have a positive safety loading ($c > \lambda E[X]$), and interest zero.

Using equicontinuity of $V'_\Delta(s)$ one can show—as in Chap. IX of Fleming and Soner (2006)—that these discretisations converge (see Edalati and Hipp 2013 and Hipp 2015; for the case of Example 1, see also Belkina et al. 2012).

Numerical experiments show that the Euler type discretisations seem to converge also in cases in which the regularity conditions for the mentioned proofs are not satisfied. In the following example the claim size distribution is purely discrete.

Example 2 We consider claims X of size 1, and $\lambda = \mu = \sigma = 1, c = 2$. The above recursions lead to the two optimal amounts invested $A(s)$ for the unconstrained case ($\mathcal{A}(s) = (-\infty, \infty)$, dashed line) and the case without leverage and short-selling ($\mathcal{A}(s) = [0, s]$, solid line). Notice that $A(1) = 0$ in both cases (Fig. 2).

3 Dividend Payment with Ruin Constraint

For the risk process, we again use a classical Lundberg model. If $D(t)$ is the accumulated dividend stream of some admissible dividend payment strategy, then the dividend value for a discount rate $\delta > 0$ is

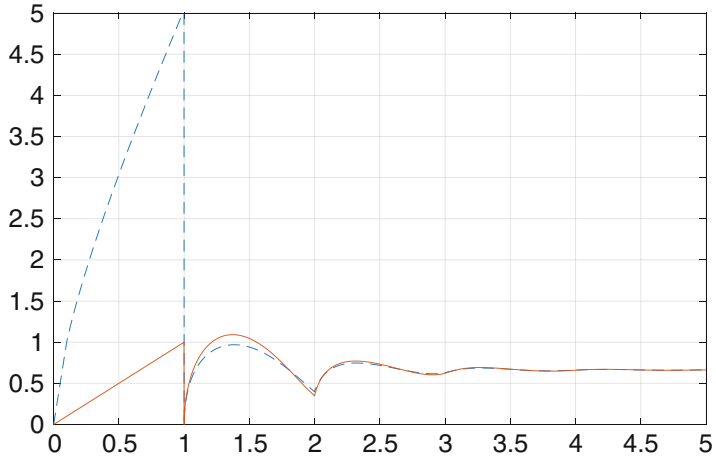


Fig. 2 Optimal amount invested $A(s)$ for $0 \le s \le 5$

$$V^D(s) = E\left[\int_0^\infty e^{-\delta t} dD(t) | S(0) = s\right].$$

Here we tacitly assume that no dividends are paid at or after ruin. The value of the company (without ruin constraint) is

$$V_0(s) = \sup_D V^D(s),$$

where the supremum is taken over all admissible dividend payment strategies. For many popular claim size distributions, the optimal dividend payment strategy is a barrier strategy (see Loeffen 2008) with barrier M, say. The computation of $V_0(s)$ and M is based on the dynamic equation

$$0 = \delta V_0(s) + \mathcal{G} V_0(s), \tag{5}$$

where

$$\mathcal{G} f(s) = \lambda E[f(s - X) - f(s)] + c f'(s)$$

is the infinitesimal generator of the Lundberg model. If $v(s)$ is the solution to (5) with $v(0) = v'(0) = 1$, then

$$M = \arg\min v'(s)$$

and

$$V_0(s) = v(s)/v'(M),\ s \le M,\ V_0(s) = V_0(M) + s - M,\ s \ge M.$$

The company value with ruin constraint is

$$V(s,\alpha) = \sup_D [V^D(s) : \psi^D(s) \le \alpha],$$

where $\psi^D(s)$ is the ruin probability of the with dividend process $S^D(t)$ for initial surplus s. The corresponding problem with Lagrange multiplier L is

$$V(s,L) = \sup_D [V^D(s) - L\psi^D(s)].$$

These two concepts are not equivalent: we have

$$V(s,L) = \sup\{V(s,\alpha) - L\alpha : 0 \le \alpha \le 1\},$$

but it might happen that we have $\alpha_1 < \alpha_2$ for which

$$V(s,L) = V(s,\alpha_1) - L\alpha_1 = V(s,\alpha_2) - L\alpha_2,$$

and then for $\alpha_1 < \alpha < \alpha_2$ we cannot find any L for which

$$V(s,L) = V(s,\alpha) - L\alpha.$$

This situation is called *Lagrange gap.*

We will compute $V(s,L)$ and the corresponding with dividend ruin probability α. This way, we also obtain $V(s,\alpha)$, at least for such values of α. The computation is based on a non-stationary approach: for time t we consider dividends payment and ruin after time t, where dividends are discounted to time 0:

$$W(s,t) = \sup_D \left[E\left[\int_t^\infty e^{-\delta u} dD(u) | S(t) = s \right] - L\mathbb{P}\{\inf_{u \ge t} S^D(u) < 0 | S(t) = s\} \right].$$

The functions $W(s,t)$ satisfy the dynamic equation

$$0 = W_t(s,t) + \mathcal{G}W(s,t), s, t \ge 0. \tag{6}$$

This is the dynamic equation (5) where the term for discounting is replaced by a term for time dependence in a non-stationary model. Then $W(s,\infty) = -L\psi(s)$ leads to the following approximation: for large T we let $W(s,T) = -L\psi(s)$, and then we calculate backward the functions $W(s,t)$ to use $W(s,0)$ as an approximation for $V(s,L)$. The following numerical example has exponential claims with mean 1, premium rate $c = 2$, discount rate $\delta = 0.03$ and claim frequency $\lambda = 1$. The calculation is based on the following recursion of discretisations:

$$W_\Delta(s, t - dt) = W_\Delta(s,t) + dt\mathcal{G}W_\Delta(s,t),$$

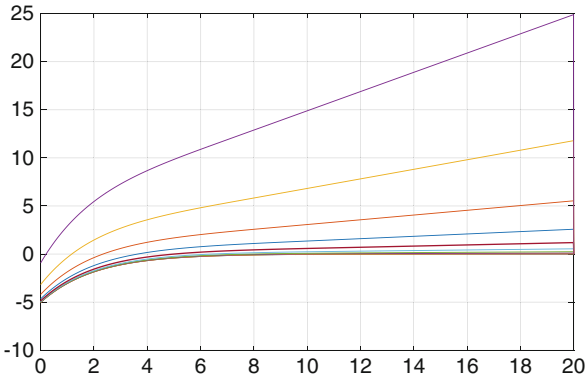


Fig. 3 The functions $W(s,t)$ for $0 \le s \le 20$

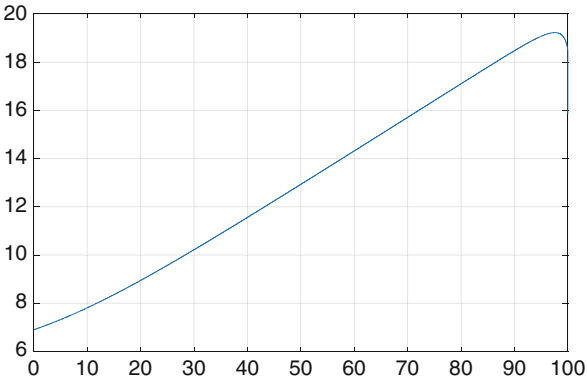


Fig. 4 The barrier $M(t)$ for $0 \le t \le 100$

where in $\mathcal{G}W_\Delta(s,t)$ we use difference ratios instead of derivatives, as in the Euler type approach above. The time dependent barrier $M(t)$ is defined as the first value s at which

$$(W_\Delta(s,t-dt) - W_\Delta(s-ds,t-dt))/\Delta < e^{-\delta t},$$

and $W_\Delta(s,t-dt)$ is linear on $[M(t),\infty)$ with slope $e^{-\delta t}$.

The following figures are calculated with $T = 100,\ ds = 0.01$ and $dt = 0.001$.

The function $M(t)$ should be increasing. It drops in Fig. 2 close to T, but this is a typical artefact caused by the definition of $W(s,T)$ (Fig. 4).

This shows that dividend values with a ruin constraint can be computed numerically. As a next step one should use this objective for the control of reinsurance and (constrained) investment.

Our last figure shows the efficiency curve for dividend values and ruin probabilities. It is computed with $s = 5,\ T = 300,\ ds = 0.01,\ dt = 0.001$ and $0 \le L \le 600$. The value without ruin constraint is $V_0(s) = 12.669$. One would conclude that there is no Lagrange gap here (Fig. 5).

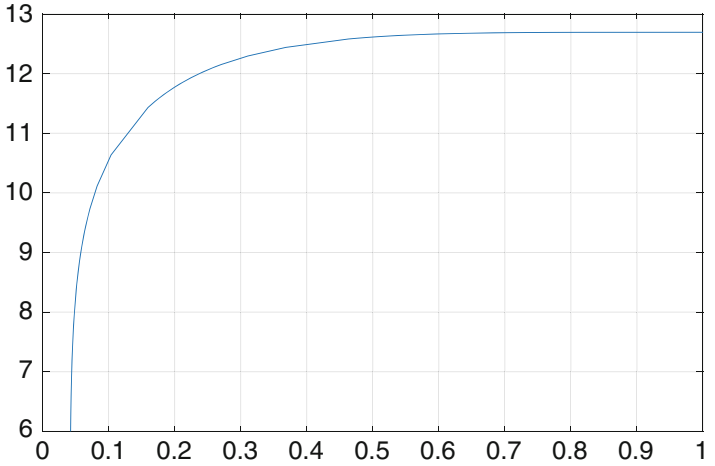


Fig. 5 $V(s, L)$ against the corresponding α-values

4 Conclusions and Future Work

For stochastic control in insurance the classical Hamilton-Jacobi-Bellman equations are still useful for infinite horizon problems; a finite horizon view is not appropriate here since insurance uses diversification in time. Viscosity solutions of these equations can be derived with simple Euler schemes, they converge under weak assumptions. This solves problems in which the ruin probability is minimized or the company value is maximized. An objective function connecting these two opposite views is a company value with a ruin constraint. For the computation of this quantity, a Lagrange approach and an appropriate discretisation are given leading to dividend strategies with a barrier which increases with time. For this given barrier, the corresponding ruin probability can be computed as well.

When the concept of a company value with ruin constraint is well understood (concerning algorithms and proof of convergence), one could and should start optimizing this objective by control of investment, reinsurance, and other control variables in insurance. If the above non-stationary approach is used, the numerics for non-linear partial differential equations will be of major importance.

References

Azcue, P., Muler, N.: Stochastic Optimization in Insurance: A Dynamic Programming Approach. Springer, New York (2014)

Belkina, T., Hipp, C., Luo, S., Taksar, M.: Optimal constrained investment in the Cramer-Lundberg model. Scand. Actuar. J. **5**, 383–404 (2014)

Black, F., Scholes, M.: The pricing of options and corporate liabilities. J. Polit. Econ. **81**(3), 637–654 (1973)
Borch, K.: The theory of risk. J. R. Stat. Soc. Ser. B **29**(3), 432–467 (1967)
Browne, S.: Optimal investment policies for a firm with a random risk process: exponential utility and minimizing the probability of ruin. Math. Oper. Res. **20**(4), 937–958 (1995)
Edalati, A., Hipp, C.: Solving a Hamilton-Jacobi-Bellman equation with constraints. Stochastics **85**(4), 637–651 (2013)
Fleming, W.H., Rishel, R.: Deterministic and Stochastic Optimal Control. Applications of Mathematics. Springer, New York (1975)
Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions. Stochastic Modelling and Applied Probability. Springer, New York (2006)
Hipp, C., Plum, M.: Optimal investment for insurers. Insur. Math. Econ. **27**(2), 215–228 (2000)
Hipp, C., Plum, M.: Optimal investment for investors with state dependent income, and for insurers. Financ. Stoch. **7**(3), 299–321 (2003)
Hipp, C.: Correction note to: solving a Hamilton-Jacobi-Bellman equation with constraints. Stochastics **88**(4), 481–490 (2015)
Kalashnikov, V., Norberg, R.: Power tailed ruin probabilities in the presence of risky investments. Stoch. Process. Appl. **98**(2), 211–228 (2002)
Karatzas, I., Shreve, S.: Methods of Mathematical Finance. Springer, New York (1998)
Loeffen, R.L.: On optimality of the barrier strategy in de Finettis dividend problem for spectrally negative Lévy processes. Ann. Appl. Probab. **18**(5), 1669–1680 (2008)
Merton, R.C.: Lifetime portfolio selection under uncertainty: the continuous-time case. Rev. Econ. Stat. **51**(3), 247–257 (1969)
Merton, R.C.: Lifetime portfolio selection under uncertainty: the continuous-time case. J. Econ. Theory **3**, 373–413 (1971)
Merton, R.C.: Theory of rational option pricing. Bell J. Econ. Manag. Sci. **4**(1), 141–183 (1973)
Merton, R.C.: Continuous Time Finance.Wiley-Blackwell, New York (1992)
Øksendal, B.: Applied Stochastis Control of Jump Diffusions. Springer, New York (2005)
Pham, H.: Continuous-Time Stochastic Control and Optimization with Financial Applications. Springer, New York (2009)
Schmidli, H.: Stochastic Control in Insurance. Springer, New York (2007)
Stein, J.L.: Stochastic Optimal Control and the U.S. Financial Debt Crisis. Springer, New York (2012)

Part II
Quantitative Finance

Bermudan Option Valuation Under State-Dependent Models

Anastasia Borovykh, Andrea Pascucci, and Cornelis W. Oosterlee

Abstract We consider a defaultable asset whose risk-neutral pricing dynamics are described by an exponential Lévy-type martingale. This class of models allows for a local volatility, local default intensity and a locally dependent Lévy measure. We present a pricing method for Bermudan options based on an analytical approximation of the characteristic function combined with the COS method. Due to a special form of the obtained characteristic function the price can be computed using a fast Fourier transform-based algorithm resulting in a fast and accurate calculation.

Keywords Bermudan option • Local Lévy model • Defaultable asset • Asymptotic expansion • Fourier-cosine expansion

1 Introduction

In order to price derivatives in finance one requires the specification of the underlying asset dynamics. This is usually done by means of a stochastic differential equation. In this work we consider the flexible dynamics of a state-dependent model, in which we account for a local volatility function, a local jump measure such that the jumps in the underlying arrive with a state-dependent intensity and a local default intensity, so that the default time depends on the underlying state. One of the problems when considering such a state-dependent model is the fact that there is no explicit density function or characteristic function available. In order to still be able to price derivatives, we derive the characteristic function by means of an advanced Taylor expansion of the state-dependent coefficients, as first presented in Pagliarani et al. (2013) for a simplified model and similar to the derivations in Borovykh

A. Borovykh (✉) • A. Pascucci
Dipartimento di Matematica, Università di Bologna, Bologna, Italy
e-mail: borovykh_a@hotmail.com; andrea.pascucci@unibo.it

C.W. Oosterlee
Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
Delft University of Technology, Delft, The Netherlands
e-mail: c.w.oosterlee@cwi.nl

J.A. Londoño et al. (eds.), *Actuarial Sciences and Quantitative Finance*, Springer
Proceedings in Mathematics & Statistics 214, https://doi.org/10.1007/978-3-319-66536-8_6

et al. (2016) for the local Lévy model. This Taylor expansion allows one to rewrite the fundamental solution of the related Cauchy problem in terms of solutions of simplified Cauchy problems, which we then solve in the Fourier space to obtain the approximated characteristic function. Once we have an explicit approximation for the characteristic function we use a Fourier method known as the COS method, first presented in Fang and Oosterlee (2009), for computing the continuation value of a Bermudan option. Due to a specific form of the approximated characteristic function the continuation value can be computed using a Fast Fourier Transform (FFT), resulting in a fast and accurate option valuation.

2 General Framework

We consider a defaultable asset S whose risk-neutral dynamics are given by:

$$\begin{aligned} S_t &= \mathbb{1}_{\{t<\zeta\}} e^{X_t}, \\ dX_t &= \mu(t, X_t)dt + \sigma(t, X_t)dW_t + \int_{\mathbb{R}} d\tilde{N}_t(t, X_{t-}, dz)z, \\ d\tilde{N}_t(t, X_{t-}, dz) &= dN_t(t, X_{t-}, dz) - \nu(t, X_{t-}, dz)dt, \\ \zeta &= \inf\{t \geq 0 : \int_0^t \gamma(s, X_s)ds \geq \varepsilon\}, \end{aligned}$$

where $\tilde{N}_t(t, x, dz)$ is a compensated random measure with state-dependent Lévy measure $\nu(t, x, dz)$. The default time ζ of S is defined in a canonical way as the first arrival time of a doubly stochastic Poisson process with local intensity function $\gamma(t, x) \geq 0$, and $\varepsilon \sim \text{Exp}(1)$ and is independent of X. Thus the model features:

- a local volatility function $\sigma(t, x)$;
- a local Lévy measure: jumps in X arrive with a state-dependent intensity described by the local Lévy measure $\nu(t, x, dz)$. The jump intensity and jump distribution can thus change depending on the value of x. A state-dependent Lévy measure is an important feature because it allows to incorporate stochastic jump-intensity into the modeling framework;
- a local default intensity $\gamma(t, x)$: the asset S can default with a state-dependent default intensity.

We define the filtration of the market observer to be $\mathcal{G} = \mathcal{F}^X \vee \mathcal{F}^D$, where $\mathcal{F}^X$ is the filtration generated by X and $\mathcal{F}_t^D := \sigma(\{\zeta \leq u\}, u \leq t)$, for $t \geq 0$, is the filtration of the default. We assume

$$\int_{\mathbb{R}} e^{|z|}\nu(t, x, dz) < \infty,$$

and by imposing that the discounted asset price $\tilde{S}_t := e^{-rt}S_t$ is a $\mathscr{G}$-martingale, we get the following restriction on the drift coefficient:

$$\mu(t,x) = \gamma(t,x) + r - \frac{\sigma^2(t,x)}{2} - \int_{\mathbb{R}} \nu(t,x,dz)(e^z - 1 - z).$$

3 The Characteristic Function

Is it well-known (see, for instance, Linetsky 2006, Sect. 2.2) that the price V of a European option with maturity T and payoff $\Phi(S_T)$ is given by

$$V_t = \mathbb{1}_{\{\zeta > t\}} e^{-r(T-t)} E\left[e^{-\int_t^T \gamma(s,X_s)ds}\varphi(X_T)|X_t\right], \qquad t \le T,$$

where $\varphi(x) = \Phi(e^x)$. Thus, in order to compute the price of an option, we must evaluate functions of the form

$$u(t,x) := E\left[e^{-\int_t^T \gamma(s,X_s)ds}\varphi(X_T)|X_t = x\right]. \tag{2}$$

Under standard assumptions, u can be expressed as the classical solution of the following Cauchy problem

$$\begin{cases} Lu(t,x) = 0, & t \in [0,T[,\ x \in \mathbb{R}, \\ u(T,x) = \varphi(x), & x \in \mathbb{R}, \end{cases} \tag{3}$$

where L is the integro-differential operator

$$\begin{aligned} Lu(t,x) &= \partial_t u(t,x) + r\partial_x u(t,x) + \gamma(t,x)(\partial_x u(t,x) - u(t,x)) \\ &\quad + \frac{\sigma^2(t,x)}{2}(\partial_{xx} - \partial_x)u(t,x) - \int_{\mathbb{R}} \nu(t,x,dz)(e^z - 1 - z)\partial_x u(t,x) \\ &\quad + \int_{\mathbb{R}} \nu(t,x,dz)(u(t,x+z) - u(t,x) - z\partial_x u(t,x)). \end{aligned} \tag{4}$$

Define $\Gamma(t,x;T,y)$ to be the fundamental solution of the Cauchy problem (3). The function u in (2) can be represented as an integral with respect to $\Gamma(t,x;T,dy)$:

$$u(t,x) = \int_{\mathbb{R}} \varphi(y)\Gamma(t,x;T,dy). \tag{5}$$

Here we notice explicitly that $\Gamma(t,x;T,dy)$ is not necessarily a standard probability measure because its integral over $\mathbb{R}$ can be strictly less than one; nevertheless, with a slight abuse of notation, we refer to its Fourier transform

$$\hat{\Gamma}(t,x;T,\xi) := \mathscr{F}(\Gamma(t,x;T,\cdot))(\xi) := \int_{\mathbb{R}} e^{i\xi y}\Gamma(t,x;T,dy), \qquad \xi \in \mathbb{R},$$

as the characteristic function of $\log S$. Following the method developed in Borovykh et al. (2016) we use an adjoint expansion of the state-dependent coefficients

$$a(t,x) := \frac{\sigma^2(t,x)}{2}, \qquad \gamma(t,x), \qquad \nu(t,x,dz),$$

around some point $\bar{x}$. The coefficients $a(t,x)$, $\gamma(t,x)$ and $\nu(t,x,dz)$ are assumed to be continuously differentiable with respect to x up to order $N \in \mathbb{N}$. Introducing the n-th order Taylor approximation of the operator L to be (4):

$$\begin{aligned} L_n = L_0 + \sum_{k=1}^{n} \Big(&(x-\bar{x})^k a_k(\partial_{xx} - \partial_x) + (x-\bar{x})^k \gamma_k \partial_x - (x-\bar{x})^k \gamma_k \\ &- \int_{\mathbb{R}} (x-\bar{x})^k \nu_k(dz)(e^z - 1 - z)\partial_x + \int_{\mathbb{R}} (x-\bar{x})^k \nu_k(dz)(e^{z\partial_x} - 1 - z\partial_x) \Big), \end{aligned}$$

where

$$\begin{aligned} L_0 = &\partial_t + r\partial_x + a_0(t)(\partial_{xx} - \partial_x) + \gamma_0(t)\partial_x - \gamma_0(t) - \int_{\mathbb{R}} \nu_0(t,dz)(e^z - 1 - z)\partial_x \\ &+ \int_{\mathbb{R}} \nu_0(t,dz)(e^{z\partial_x} - 1 - z\partial_x), \end{aligned}$$

and

$$a_k = \frac{\partial_x^k a(\bar{x})}{k!}, \qquad \gamma_k = \frac{\partial_x^k \gamma(\bar{x})}{k!}, \qquad \nu_k(dz) = \frac{\partial_x^k \nu(\bar{x},dz)}{k!}, \qquad k \geq 0.$$

Let us assume for a moment that L_0 has a fundamental solution $G^0(t,x;T,y)$ that is defined as the solution of the Cauchy problem

$$\begin{cases} L_0 G^0(t,x;T,y) = 0 \quad t \in [0,T[,\ x \in \mathbb{R}, \\ G^0(T,\cdot;T,y) = \delta_y. \end{cases}$$

In this case we define the nth-order approximation of Γ as

$$\Gamma^{(n)}(t,x;T,y) = \sum_{k=0}^{n} G^k(t,x;T,y),$$

where, for any $k \geq 1$ and (T,y), $G^k(\cdot,\cdot;T,y)$ is defined recursively through the following Cauchy problem

$$\begin{cases} L_0 G^k(t,x;T,y) = -\sum_{h=1}^{k} (L_h - L_{h-1}) G^{k-h}(t,x;T,y) & t \in [0,T[,\ x \in \mathbb{R}, \\ G^k(T,x;T,y) = 0, & x \in \mathbb{R}. \end{cases}$$

Correspondingly, the nth-order approximation of $\hat{\Gamma}$ is defined to be

$$\hat{\Gamma}^{(n)}(t,x;T,\xi) = \sum_{k=0}^{n} \mathscr{F}\left(G^k(t,x;T,\cdot)\right)(\xi) := \sum_{k=0}^{n} \hat{G}^k(t,x;T,\xi), \qquad \xi \in \mathbb{R}.$$

Now, by transforming the simplified Cauchy problems into adjoint problems and solving these in the Fourier space we find

$$\hat{G}^0(t,x;T,\xi) = e^{i\xi x} e^{\int_t^T \psi(s,\xi)ds},$$

$$\hat{G}^k(t,x;T,\xi) = -\int_t^T e^{\int_s^T \psi(\tau,\xi)d\tau} \mathscr{F}\left(\sum_{h=1}^{k} \left(\tilde{L}_h^{(s,\cdot)}(s) - \tilde{L}_{h-1}^{(s,\cdot)}(s)\right) G^{k-h}(t,x;s,\cdot)\right)(\xi)ds,$$

with

$$\begin{aligned} \psi(s,\xi) =& i\xi(r + \gamma_0(s)) + a_0(s)(-\xi^2 - i\xi) - \int_{\mathbb{R}} \nu_0(s,dz)(e^z - 1 - z)i\xi \\ &+ \int_{\mathbb{R}} \nu_0(s,dz)(e^{iz\xi} - 1 - iz\xi), \end{aligned}$$

the characteristic exponent of the Lévy process with coefficients $\gamma_0(s)$, $a_0(s)$ and $\nu_0(s,dz)$, and

$$\begin{aligned} \tilde{L}_h^{(s,y)}(s) - \tilde{L}_{h-1}^{(s,y)}(s) =& a_h(s)h(h-1)(y-\bar{x})^{h-2} \\ &+ a_h(s)(y-\bar{x})^{h-1}\left(2h\partial_y + (y-\bar{x})(\partial_{yy} + \partial_y) + h\right) \\ &- \gamma_h(s)h(y-\bar{x})^{h-1} - \gamma_h(s)(y-\bar{x})^h\left(\partial_y + 1\right) \\ &+ \int_{\mathbb{R}} \nu_h(s,dz)(e^z - 1 - z)\left(h(y-\bar{x})^{h-1} + (y-\bar{x})^h\partial_y\right) \\ &+ \int_{\mathbb{R}} \bar{\nu}_h(s,dz)\big((y+z-\bar{x})^h e^{z\partial_y} \\ &- (y-\bar{x})^h - z\left(h(y-\bar{x})^{h-1} - (y-\bar{x})^h\partial_y\right)\big). \end{aligned}$$

From these results one can already see that the dependency on x comes in through $e^{i\xi x}$ and after taking derivatives the dependency on x will take the form $(x-\bar{x})^m e^{i\xi x}$: this fact will be crucial in our analysis. After some algebraic manipulations, see

for details Borovykh et al. (2016), we find that the approximation of order n is a function of the form

$$\hat{\Gamma}^{(n)}(t,x;T,\xi) := e^{i\xi x}\sum_{k=0}^{n}(x-\bar{x})^k g_{n,k}(t,T,\xi), \tag{6}$$

where the coefficients $g_{n,k}$, with $0 \leq k \leq n$, depend only on t, T and ξ, but not on x. The approximation formula can thus always be split into a sum of products of functions depending only on ξ and functions that are linear combinations of $(x-\bar{x})^m e^{i\xi x}$, $m \in \mathbb{N}_0$.

4 Bermudan Option Valuation

A Bermudan option is a financial contract in which the holder can exercise at a predetermined finite set of exercise moments prior to maturity, and the holder of the option receives a payoff when exercising. Consider a Bermudan option with a set of M exercise moments $\{t_1,\ldots,t_M\}$, with $0 \leq t_1 < t_2 < \cdots < t_M = T$. When the option is exercised at time t_m the holder receives the payoff $\Phi(t_m, S_{t_m})$. For a Bermudan put option with strike price K, we simply have $\varphi(t,x) = (K-e^x)^+$. By the dynamic programming approach, the option value can be expressed by a backward recursion as

$$v(t_M,x) = \mathbb{1}_{\{\zeta>t_M\}}\varphi(t_M,x)$$

and

$$\begin{cases} c(t,x) = E\left[e^{\int_t^{t_m}-(r+\gamma(s,X_s))ds}v(t_m,X_{t_m})|X_t=x\right], & t\in[t_{m-1},t_m[\\ v(t_{m-1},x) = \mathbb{1}_{\{\zeta>t_{m-1}\}}\max\{\varphi(t_{m-1},x),c(t_{m-1},x)\}, & m\in\{2,\ldots,M\}. \end{cases} \tag{7}$$

In the above notation $v(t,x)$ is the option value and $c(t,x)$ is the so-called continuation value. The option value is set to be $v(t,x) = c(t,x)$ for $t \in]t_{m-1},t_m[$, and, if $t_1 > 0$, also for $t \in [0,t_1[$.

Remark 4.1 Since the payoff of a call option grows exponentially with the log-stock price, this may introduce significant cancellation errors for large domain sizes. For this reason we price put options only using our approach and we employ the well-known put-call parity to price calls via puts. This is a rather standard argument (see, for instance, Zhang and Oosterlee 2012).

4.1 An Algorithm for Pricing Bermudan Put Options

The COS method as proposed in Fang and Oosterlee (2009) is based on the insight that the Fourier-cosine series coefficients of $\Gamma(t,x;T,dy)$ (and therefore also of option prices) are closely related to the characteristic function of the underlying process. Remembering that the expected value $c(t,x)$ in (7) can be rewritten in integral form as in (5),

$$c(t,x) = e^{-r(t_m-t)} \int_{\mathbb{R}} v(t_m,y)\Gamma(t,x;t_m,dy), \qquad t \in [t_{m-1},t_m[,$$

we apply the COS formulas to find the approximation:

$$\hat{c}(t,x) = e^{-r(t_m-t)} \sum_{k=0}^{N-1}{}' \operatorname{Re}\left(e^{-ik\pi\frac{a}{b-a}}\hat{\Gamma}\left(t,x;t_m,\frac{k\pi}{b-a}\right)\right) V_k(t_m), \qquad t \in [t_{m-1},t_m[\tag{8}$$

$$V_k(t_m) = \frac{2}{b-a}\int_a^b \cos\left(k\pi\frac{y-a}{b-a}\right)\max\{\varphi(t_m,y),c(t_m,y)\}dy,$$

with $\varphi(t,x) = (K-e^x)^+$.

Next we recover the coefficients $(V_k(t_m))_{k=0,1,\ldots,N-1}$ from $(V_k(t_{m+1}))_{k=0,1,\ldots,N-1}$. To this end, we split the integral in the definition of $V_k(t_m)$ into two parts using the early-exercise point x_m^*, which is the point where the continuation value is equal to the payoff, i.e. $c(t_m,x_m^*) = \varphi(t_m,x_m^*)$; thus, we have

$$V_k(t_m) = F_k(t_m,x_m^*) + C_k(t_m,x_m^*), \qquad m = M-1,M-2,\ldots,1,$$

where

$$F_k(t_m,x_m^*) := \frac{2}{b-a}\int_a^{x_m^*} \varphi(t_m,y)\cos\left(k\pi\frac{y-a}{b-a}\right)dy,$$

$$C_k(t_m,x_m^*) := \frac{2}{b-a}\int_{x_m^*}^b c(t_m,y)\cos\left(k\pi\frac{y-a}{b-a}\right)dy,$$

and $V_k(t_M) = F_k(t_M,\log K)$.

Remark 4.2 Since we have a semi-analytic formula for $\hat{c}(t_m,x)$, we can easily find the derivatives with respect to x and use Newton's method to find the point x_m^* such that $c(t_m,x_m^*) = \varphi(t_m,x_m^*)$. A good starting point for the Newton method is $\log K$, since $x_m^* \le \log K$.

The coefficients $F_k(t_m, x_m^*)$ can be computed analytically using $x_m^* \leq \log K$. On the other hand, by inserting the approximation (8) for the continuation value into the formula for $C_k(t_m, x_m^*)$ have the following coefficients $\hat{C}_k$ for $m = M-1, M-2, \ldots, 1$:

$$\hat{C}_k(t_m, x_m^*) = \frac{2e^{-r(t_{m+1}-t_m)}}{b-a} \sum_{j=0}^{N-1}{}' V_j(t_{m+1}) \int_{x_m^*}^{b} \text{Re}\left(e^{-ij\pi \frac{a}{b-a}} \hat{\Gamma}\left(t_m, x; t_{m+1}, \frac{j\pi}{b-a}\right)\right) \cos\left(k\pi \frac{x-a}{b-a}\right) dx.$$

Similar to the FFT-based algorithm in Fang and Oosterlee (2009) for an exponential Lévy process with constant coefficients, the continuation value in case of the state-dependent coefficients can also be calculated using the FFT. Using the structure of the characteristic function (6) we write the continuation value in vector form as:

$$\hat{\mathbf{C}}(t_m, x_m^*) = \sum_{h=0}^{n} e^{-r(t_{m+1}-t_m)} \text{Re}\left(\mathbf{V}(t_{m+1}) \mathscr{M}^h(x_m^*, b) \Lambda^h\right),$$

where $\mathbf{V}(t_{m+1})$ is the vector $[V_0(t_{m+1}), \ldots, V_{N-1}(t_{m+1})]^T$ and $\mathscr{M}^h(x_m^*, b)\Lambda^h$ is a matrix-matrix product with $\mathscr{M}^h$ being a matrix with elements

$$M_{k,j}^h(x_m^*, b) = \frac{2}{b-a} \int_{x_m^*}^{b} e^{ij\pi \frac{x-a}{b-a}} (x-\bar{x})^h \cos\left(k\pi \frac{x-a}{b-a}\right) dx, \qquad k, j = 0, \ldots, N-1$$

and Λ^h is a diagonal matrix with elements

$$g_{n,h}\left(t_m, t_{m+1}, \frac{j\pi}{b-a}\right), \qquad j = 0, \ldots, N-1.$$

It can be shown using standard trigonometric that the matrix $\mathscr{M}$ can be rewritten as a sum of a Hankel and Toeplitz matrix such that $\mathscr{M} = \mathscr{M}_H + \mathscr{M}_T$ with elements

$$M_j^h(x_m^*, b) = \frac{1}{b-a} \int_{x_m^*}^{b} \cos\left(ij\pi \frac{x-a}{b-a}\right) (x-\bar{x})^h dx + \frac{1}{b-a} \int_{x_m^*}^{b} \sin\left(ij\pi \frac{x-a}{b-a}\right) (x-\bar{x})^h dx.$$

Using the split into sums of Hankel and Toeplitz matrices we can write the continuation value in matrix form as:

$$\hat{\boldsymbol{C}}(t_m, x_m^*) = \sum_{h=0}^{n} e^{-r(t_{m+1}-t_m)} \text{Re}\left((\mathscr{M}_H^h + \mathscr{M}_T^h)\boldsymbol{u}^h\right),$$

where $\mathscr{M}_H^h = \{M_{k,j}^{H,h}(x_m^*, b)\}_{k,j=0}^{N-1}$ is a Hankel matrix and $\mathscr{M}_T^l = \{M_{k,j}^{T,h}(x_m^*, b)\}_{k,j=0}^{N-1}$ is a Toeplitz matrix and $\boldsymbol{u}^h = \{u_j^h\}_{j=0}^{N-1}$, with $u_j^h = g_{n,h}\left(t_m, t_{m+1}, \frac{j\pi}{b-a}\right) V_j(t_{m+1})$ and $u_0^h = \frac{1}{2} g_{n,h}(t_m, t_{m+1}, 0)\, V_0(t_{m+1})$. It is well-known that a product of a Hankel or Toeplitz matrix with a vector can be calculated using FFTs, see Borovykh et al. (2016) for full details. Using the fact that an FFT can be computed with computational complexity $O(N \log_2 N)$, we find that for a Bermudan option with M exercise dates the overall computational complexity is $O((M-1)N \log_2 N)$.

5 Numerical Experiments

In this section we apply the method developed in Sect. 4 to compute the European and Bermudan option values with various underlying stock dynamics. The computer used in the experiments has an Intel Core i7 CPU with a 2.2 GHz processor. We use the second-order approximation of the characteristic function.

For the COS method, unless otherwise mentioned, we use $N = 200$ and $L = 10$, where L is the parameter used to define the truncation range $[a, b]$ as follows:

$$[a, b] := \left[c_1 - L\sqrt{c_2 + \sqrt{c_4}}, c_1 + L\sqrt{c_2 + \sqrt{c_4}}\right],$$

where c_n is the nth cumulant of log-price process $\log S$ calculated using the 0th-order approximation of the characteristic function. We compare the approximated values to a 95% confidence interval computed with a Longstaff-Schwartz method with 10^5 simulations and 250 time steps per year. Furthermore, in the expansion we always use $\bar{x} = X_0$.

5.1 Tests Under CEV-Merton Dynamics

Consider a process under the CEV-Merton dynamics:

$$dX_t = \left(r - a(X_t) - \lambda\left(e^{m+\delta^2/2} - 1\right)\right) dt + \sqrt{2a(X_t)} dW_t + \int_{\mathbb{R}} d\tilde{N}_t(t, dz) z,$$

with

$$a(x) = \frac{\sigma_0^2 e^{2(\beta-1)x}}{2},$$

$$\nu(dz) = \lambda \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(\frac{-(z-m)^2}{2\delta^2}\right) dz,$$

$$\psi(\xi) = -a_0(\xi^2 + i\xi) + ir\xi - i\lambda\left(e^{m+\delta^2/2} - 1\right)\xi + \lambda\left(e^{mi\xi - \delta^2\xi^2/2} - 1\right).$$

Table 1 Prices for a European and a Bermudan put option (expiry $T = 1$ with 10 exercise dates and expiry $T = 2$ with 20 exercise dates) in the CEV-Merton model for the 2nd-order approximation of the characteristic function, and a Monte Carlo method

		European		Bermudan	
T	K	MC 95% c.i.	Value	MC 95% c.i.	Value
1	0.6	0.006136–0.006573	0.006579	0.006307–0.006729	0.006096
	0.8	0.02526–0.02622	0.02581	0.02595–0.2689	0.02520
	1	0.08225–0.08395	0.08250	0.08480–0.08640	0.08593
	1.2	0.1965–0.1989	0.1977	0.2097–0.2115	0.2132
	1.4	0.3560–0.3589	0.3574	0.3946–0.3957	0.3954
	1.6	0.5341–0.5385	0.5364	0.5930–0.5941	0.5932
2	0.6	0.01444–0.01513	0.01529	0.01528–0.01594	0.01365
	0.8	0.04522–0.04655	0.04613	0.04596–0.04719	0.04659
	1	0.1046–0.1067	0.1077	0.1149–0.1170	0.1171
	1.2	0.2054–0.2083	0.2065	0.2319–0.2345	0.2345
	1.4	0.3351–0.3386	0.3382	0.3968–0.3987	0.3991
	1.6	0.4904–0.4944	0.4919	0.5927–0.5938	0.5935

We use the following parameters $S_0 = 1$, $r = 5\%$, $\sigma_0 = 20\%$, $\beta = 0.5$, $\lambda = 30\%$, $m = -10\%$, $\delta = 40\%$ and compute the European and Bermudan option values in Table 1. The results are compared to a widely used method for valuing Bermudan options, the Least-Squares Monte Carlo method (LSM), see Longstaff and Schwartz (2001). The error in our approximation consists of the error of the COS method and the error in the adjoint expansion of the characteristic function. In particular for low strikes the method seems to be more sensitive to the approximation, as the approximated value does not always fall into the LSM confidence interval.

In Fig. 1 the convergence results of the COS method using the 2nd-order approximation of the characteristic function for $T = 1$ and 10 exercise dates are presented. We choose $L = 10$ and $N = 2^d$ and see that a very quick convergence is obtained.

5.2 *Tests Under a CEV-Like Lévy Process with a State-Dependent Measure*

In this section we consider a model similar to the one used in Jacquier and Lorig (2013). The model is defined with local volatility and a state-dependent Lévy measure as follows:

$$\begin{aligned} a(x) &= \frac{1}{2}(b_0^2 + \varepsilon_1 b_1^2 \eta(x)), \\ \nu(x, dz) &= \varepsilon_3 \nu_N(dz) + \varepsilon_4 \eta(x) \nu_N(dz), \\ \eta(x) &= e^{\beta x}. \end{aligned} \tag{9}$$

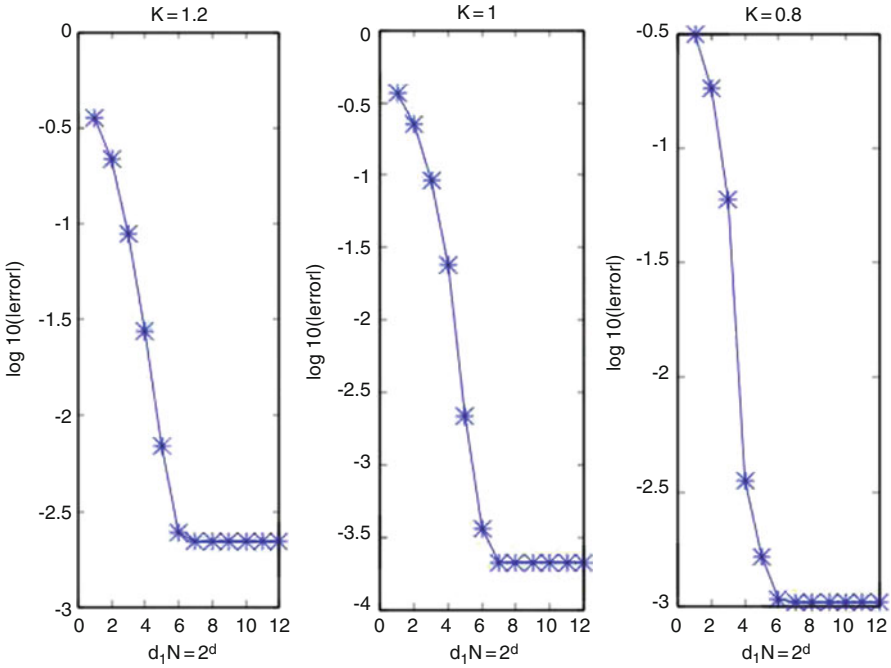


Fig. 1 Error convergence for pricing Bermudan put options, $N = 2^d$, $L = 10$, $T = 1$ and 10 exercise dates and strikes $K = 0.8, 1, 1.2$

We will consider Gaussian jumps, meaning that

$$\nu_N(dz) = \lambda \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(\frac{-(z-m)^2}{2\delta^2}\right) dz.$$

In Table 2 the results are presented for a model as defined in (9) with a state-dependent jump measure, so $\nu(x, dz) = \eta(x)\nu_N(dz)$. In this case we have

$$\psi(\xi) = ir\xi - a_0(\xi^2 - i\xi) - \lambda\nu_0(e^{m+\delta^2/2} - 1)i\xi + \lambda\nu_0(e^{mi\xi - \delta^2\xi^2/2} - 1),$$

where $a_0 = \frac{1}{2}b_1^2 e^{\beta\bar{x}}$ and $\nu_0(dz) = e^{\beta\bar{x}}\nu_N(dz)$. The other parameters are chosen as: $b_1 = 0.15$, $b_0 = 0$, $\beta = -2$, $\lambda = 20\%$, $\delta = 20\%$, $m = -0.2$, $S_0 = 1$, $r = 5\%$, $\varepsilon_1 = 1$, $\varepsilon_3 = 0$, $\varepsilon_4 = 1$, the number of exercise dates is 10 and $T = 1$. Again the method performs accurately, but for out-of- and at-the money strikes the approximation tends to under- and over-estimate the LSM value.

Table 2 Prices for a European and a Bermudan put option (10 exercise dates, expiry $T = 1$) in the CEV-like model with state-dependent measure for the 2nd-order approximation characteristic function, and a Monte Carlo method

	European		Bermudan	
K	MC 95% c.i.	Value	MC 95% c.i.	Value
0.8	0.01025–0.01086	0.009385	0.01068–0.01125	0.01024
1	0.04625–0.04745	0.04817	0.05141–0.05253	0.05488
1.2	0.1563–0.1582	0.1564	0.1942–0.1952	0.1952
1.4	0.3313–0.3334	0.3314	0.3927–0.3934	0.3930
1.6	0.5207–0.5229	0.5218	0.5919–0.5926	0.5920
1.8	0.7103–0.7124	0.7122	0.7906–0.7913	0.7910

References

Borovykh, A., Pascucci, A., Oosterlee, C.W.: Pricing Bermudan options under local Lévy models with default. J. Math. Anal. Appl. (2016)

Fang, F., Oosterlee, C.W.: Pricing early-exercise and discrete barrier options by Fourier-cosine series expansions. Numer. Math. **114**, 27–62 (2009)

Jacquier, A., Lorig, M.: The smile of certain Lévy-type models. SIAM J. Financ. Math. **4**, 804–830 (2013)

Linetsky, V.: Pricing equity derivatives subject to bankruptcy. Math. Finance **16**, 255–282 (2006)

Longstaff, F., Schwartz, E.: Valuing American Options by Simulation: A Simple Least-Squares Approach. Rev. Financ. Stud. **14**, 113–147 (2001)

Pagliarani, S., Pascucci, A., Riga, C.: Adjoint expansions in local Lévy models. SIAM J. Financ. Math. **4**, 265–296 (2013)

Zhang, B., Oosterlee, C.W.: Fourier cosine expansions and put-call relations for Bermudan options. In: Numerical Methods in Finance, vol. 12 of Springer Proc. Math., pp. 323–350. Springer, Heidelberg (2012)

Option-Implied Objective Measures of Market Risk with Leverage

Matthias Leiss and Heinrich H. Nax

Abstract Leverage has been shown to be procyclical and indicative of financial market risk. Here, we present a novel, inherently forward-looking way to estimate market leverage ratios based on derivative prices, option hedging, and the 'operational' riskiness measure by Foster and Hart (J Polit Econ 117(5):785–814, 2009). Furthermore, we report option-implied 'optimal' leverage levels inferred via the (Kelly, IRE Trans. Inf. Theory 2(3):185–189, 1956) criterion. The resulting measure of leverage exhibits strong procyclicality prior to the Global Financial Crisis of 2008. Finally, we find it to successfully predict large stock market downturns.

Keywords Objective risk • Foster-Hart • Leverage • Risk-neutral densities

1 Introduction

> With the benefit of hindsight, we clearly should have put even greater emphasis on the risks of excessive leverage.
> Hildebrand (2008)

The Global Financial Crisis of 2008 brought questions related to excessive leverage back on the table of risk regulation. Previous risk regulation frameworks (e.g., Basel I and II) posed capital requirements that were (at least partially) based on the relative riskiness of various types of assets (Hildebrand, 2008). While such risk-based capital measures signaled high stability of banks prior to the Global Financial Crisis, simple leverage ratio assessments exposed the largely undercapitalized situation of key financial actors which exacerbated the crisis. As a reaction to the crisis, the new regulatory framework (Basel III) contains a simple, non-risk-based leverage ratio requirement (Basel Committee on Banking Supervision, 2010).

Nevertheless, as Schularick and Taylor (2012) have noted, we have entered an age of unprecedented financial risk due to leverage. In particular, the vast expansion of credit and financial innovation, combined with implicit government insurance and

M. Leiss • H.H. Nax (✉)
Department of Humanities, Social and Political Sciences, ETH Zurich, Clausiusstrasse 50, 8092 Zurich, Switzerland
e-mail: mleiss@ethz.ch; hnax@ethz.ch

J.A. Londoño et al. (eds.), *Actuarial Sciences and Quantitative Finance*, Springer Proceedings in Mathematics & Statistics 214, https://doi.org/10.1007/978-3-319-66536-8_7

the prospect of rescue operations, have resulted in massively increased leverage. As a result, the financial system has become more vulnerable to endogenously generated instabilities as manifested by recurring booms and busts (Von der Becke and Sornette, 2014).

A key issue inherent to leverage is procyclicality, which means that leverage ratios are only a partial remedy. In theory, standard portfolio rules would seem to imply anticyclical leverage; high leverage when the risk premium is high. Empirically, however, procyclicality of leverage has been documented extensively (Adrian and Shin, 2014). This empirical phenomenon has been explained through increased collateral requirements during downturns creating leverage cycles (Geanakoplos, 2010): increased uncertainty and volatility of asset returns lead lenders to require tighter margins, which, in turn, mechanically implies falling prices and consequently large losses for the most leveraged investors. Importantly, both of these elements feed back on each other, thus starting the leverage cycle. Any institution in the financial system where investors hold long-term, illiquid assets that are financed by short-term liabilities is particularly at risk of this, and falling leverage can consequently lead to 'runs' on such institutions (Adrian and Shin, 2014). Perhaps serving as the most famous example, the Global Financial Crisis of 2008 started as a run on the sale and repurchase (repo) market (Gorton and Metrick, 2012).

Generally, due to procyclicality, leveraged financial markets exhibit fat tails of the return distribution and clustered volatility (Thurner et al., 2012). This suggests the use of leverage ratios as indicators for the likelihood of future financial crashes and crises. Indeed, changes in dealer repos can be used to successfully forecast changes in financial market risk as measured by the Chicago Board Options Exchange Volatility Index (VIX) index (Adrian and Shin, 2010). Similarly, intermediary leverage has been shown to be negatively aligned with the banks' Value-at-Risk (VaR) (Adrian and Shin, 2014).

Our present paper pursues a similar goal, namely to use leverage procyclicality to predict market risk. Our contribution to the existing literature is the construction of leverage ratios from derivative markets. Prior work had either focused on leverage as the ratio of collateral values to the down payment (with data generally being inaccessible, Geanakoplos 2010), or as the ratio of total assets to book equity (Adrian and Shin, 2010, 2014). By contrast, our approach will be to construct forward-looking estimates of leverage ratios based on prices of financial options. Specifically, we will use risk-neutral probability distributions to evaluate the estimated, forward-looking performance of hedged portfolios as quantified by the recently proposed 'operational' riskiness measure of Foster and Hart (2009). In our generalization of the measure, allowing leverage, the measure indicates the level of leverage at which the estimated growth rate becomes negative. We note that this is fundamentally different from previous theoretical work on optimal trading with leverage. For example, the previous study by Grossman and Vila (1992) establishes optimal dynamic trading rules subject to a leverage constraint that is given. Here, our goal is to empirically determine such a constraint in the first place.

Our findings are twofold. First, leverage ratios as constructed from derivative prices exhibit a pronounced and persistent peak prior to the Global Financial Crisis of 2008, thus quantifying the procyclical leverage regime of the market. Second,

leverage ratios are found to be indicative of extreme future market-downturns. These findings complement our own investigation of option-implied operational market risks (Leiss and Nax, 2015), particularly during the build-up of the Global Financial Crisis of 2008, where our previous, leverage-free approach had only limited reach.

2 Operational Metrics of Disaster Risk

Well-known tail measures, like Value at Risk (VaR) and Expected Shortfall (ES), have become industry standards for assessing extreme market risks (Embrechts et al., 2005). By construction, they only characterize the risk of negative events while ignoring the potential upside. On the other hand, measures of dispersion such as volatility/variance or interquartile range account for up- and downturns, but are largely blind to rare extreme events on both sides of the spectrum. For example, the widely used Sharpe ratio (Sharpe, 1994) only accounts for the first two moments of the underlying return distribution, thus implicitly (and falsely) assuming that higher moments do not matter.

Two novel measures of riskiness (by Aumann and Serrano (2008) and Foster and Hart (2009)) promise to balance both, sensitivity to extreme risks and potential gains. Formally, these measures are defined for any gamble g in the set of gambles $\mathscr{G}$ characterized by random variables with positive expectation and positive probability of negative outcomes. For any gamble $g \in \mathscr{G}$, Foster and Hart (2009) uniquely define their risk measure, FH, as the zero of[1]

$$\mathbb{E}\left[\log(1 + FH(g)g)\right] = 0, \tag{1}$$

whereas Aumann and Serrano (2008) define their risk measure, AS, as the zero of

$$\mathbb{E}\left[\exp(-AS(g)g)\right] = 1. \tag{2}$$

One issue with expression (1), which will become extremely relevant for our leverage analysis, is that, for some continuous gambles $g \in \mathscr{G}$, FH thus defined may have no positive solution. In this case, Riedel and Hellmann (2015) extend the definition consistently by setting FH to the maximum possible loss incurred by that gamble. In particular, if g is a return distribution with maximum loss of 100%, FH is bound by 1.

Importantly, definitions (1) and (2) involve forming the expectation over the whole distribution of the gamble's outcomes. Thus, FH and AS are able to capture all moments of a gamble. This is formalized by Kadan and Liu (2014), who prove that higher moments do not necessarily have a weaker effect on FH and AS. In practice,

[1]The logarithmic growth rate had entered risk analysis already earlier. Examples involve the Kelly (1956) criterion (which aims to maximize growth rate), or, very similar to Foster and Hart (2009), Whitworth (1870, p. 217).

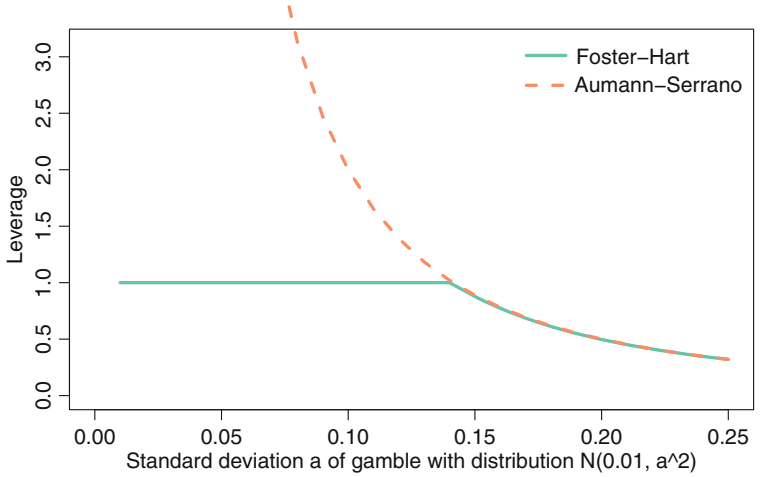


Fig. 1 Foster-Hart $FH(g)$ and Aumann-Serrano $AS(g)$ measures of riskiness vs. the standard deviation α of a normally distributed gamble $g \sim \mathcal{N}(0.01, \alpha^2)$. The implied leverage ratios coincide in the case of high risk ($\alpha \gg 0.01$). In the opposite case of vanishing risk ($\alpha \to 0$), AS diverges indicating zero risk and suggests infinite leverage, while the no-bankruptcy property of $FH(g)$ leads to an upper bound of 1

one often finds higher moments to have a strong impact on the risk measures (Kadan and Liu, 2014; Leiss and Nax, 2015; Anand et al., 2016). However, FH is significantly more sensitive to left-tail events than AS. Be g_α the composite gamble of $g_0 \in \mathcal{G}$ and an extreme loss $-L < 0$ with respective probabilities $1 - \alpha$ and $\alpha \in (0, 1)$ and $FH(g_0) > 1/L$. It is easy to show that (Kadan and Liu, 2014)

$$\lim_{\alpha \to 0} FH(g_\alpha) = 1/L, \tag{3}$$

whereas

$$\lim_{\alpha \to 0} AS(g_\alpha) = AS(g_0). \tag{4}$$

A variation of this is illustrated in Fig. 1. The gamble g is normally distributed with positive mean and standard deviation α, $g \sim \mathcal{N}(0.01, \alpha^2)$. In the high-risk scenario of large variance, $\alpha \gg 0.01$, FH and AS coincide almost perfectly. However, in the case of low risk, i.e. as $\alpha \to 0$, AS diverges indicating asymptotically zero risk and therefore infinite leverage, whereas FH is bounded by 1 to avoid bankruptcy with one shot.

Besides the above-mentioned practical appeal of taking into account the whole distribution of a gamble, both FH and AS also fill an important theoretical gap. It is known that risk-averse investors who choose their investments by maximizing expected utility may rank investments by second-order stochastic dominance

(SOSD) (Hadar and Russell, 1969; Hanoch and Levy, 1969; Rothschild and Stiglitz, 1970). However, some pairs of investments cannot be ranked on the basis of SOSD. Kadan and Liu (2014) show that both *FH* and *AS* extend SOSD in a natural way as they induce a complete ranking on $\mathscr{G}$ that agrees with SOSD whenever applicable. The induced rankings differ, because loosely speaking *FH* and *AH* order independently of an investor's utility and wealth, respectively.

The theoretical reason for *FH* to be bounded is the no-bankruptcy theorem by Foster and Hart (2009). It states that when confronted with an infinite series of gambles $g_t \in \mathscr{G}$, the simple strategy of always investing a fraction of wealth smaller than $FH(g_t)$ guarantees no-bankruptcy, i.e.

$$\mathbb{P}\left[\lim_{t\to\infty} W_t = 0\right] = 0, \tag{5}$$

where W_t denotes wealth at time t. This bound is independent of the investor's risk attitudes, which is the sense in which *FH* is 'operational' according to Foster and Hart (2009). By contrast, following such a strategy leads to wealth divergence to infinity (a.s.).

3 Extending Operational Riskiness Measures to Leveraged Gambles

The hard bound of *FH* that is induced by the no-bankruptcy constraint poses a challenge for dynamic risk management, as in some scenarios there is no more variation in *FH*. Indeed, our empirical study of option-implied *FH* found *FH* to be at the upper bound on 27% of the business days during the decade 2003–2013, and on 45% of the business days during the 5 years leading up to the collapse of Lehman Brothers in September 2008 (Leiss and Nax, 2015). One might wonder, therefore, how much information is lost because of a lack of variation during those days.

Instead of focusing on other risk indicators, we would like to explore a different 'leverage route' in this paper. Since the hard bound of one inherent to the original *FH* measure is induced by the maximal loss, one could think of building a portfolio that is hedged against extreme events: let r_s be a gamble that describes the relative return distribution of buying at asset S at time $t = 0$ and holding it until time $t = T$. Accounting for dividends paid during that period Y and discounting

$$r_s = \frac{S_T + Y - S_0}{S_0}. \tag{6}$$

If the asset defaults and no dividends are being paid, the investor incurs a maximum loss of $\min(r_s) = -100\%$ such that $FH(r_s) \leq 1$. A simple way of hedging this portfolio is via a put option written on S with premium P_0 (at $t = 0$), strike price K,

and maturity T. The return of a portfolio that consists of one unit of the stock and a put option is given by

$$r_h = \frac{\max(S_T, K) + Y - S_0 - P_0}{S_0 + P_0}, \tag{7}$$

with maximum loss of

$$\min(r_h) = \frac{K - S_0 - P_0}{S_0 + P_0} > -100\% \tag{8}$$

for $Y = 0$ and $K > 0$ (provided the seller of the option does not default). In other words, a gamble of the form (7) generally allows for $FH(r_h) > 1$, i.e. leverage.[2] Our definition (7) generalizes FH to allow for leverage.

In later sections, we will compute and analyze our 'leverage Foster-Hart' $FH(r_h)$ for hedged portfolios based on risk-neutral probability distributions estimated from option prices. Thus, the forward-looking information contained in derivative prices enter $FH(r_h)$ twice: in P_0 via the return (7), and in the computation of the expectation via (1). Figure 2 illustrates this with an example showing the payoff for investment strategy (7) for buying the S&P 500 with the corresponding put option. Here, the values are $t_0 = 2004\text{-}11\text{-}22$, $T = 2004\text{-}12\text{-}18$, $S_0 = 1177.24$ USD, $K = 1190$, $P_0 = 21.50$ USD. Note that the strike of the put is higher than index price at time $t = 0$. Option pricing according to Black and Scholes (1973) suggests that the put option ask implies a volatility of only 11.9%. In this example, one finds $FH(r_h) = 10.7$, i.e. a leverage ratio of more than 10 (see Fig. 3).

Another sensible and closely related leverage ratio is the option-implied Kelly (1956) criterion K: instead of setting the expected logarithmic growth rate to zero as in (1), one asks for that multiple (or fraction) of wealth that maximizes it, thus defining

$$\alpha_K(g) = \arg\max_{\alpha} \mathbb{E}\left[\log(1 + \alpha g)\right]. \tag{9}$$

For gambles $g \in \mathscr{G}$, one has $\alpha_K(g) \leq FH(g)$. Continuing the example from above, we obtain a maximal growth rate at a leverage ratio of $\alpha_K(r_h) = 5.1$ (see Fig. 3). The leverage ratio implied by derivative prices is not meant to be identical to other definitions (Geanakoplos, 2010; Adrian and Shin, 2010, 2014), but should be seen as complementary.

[2] Sircar and Papanicolaou (1998) document that dynamic option hedging strategies imply feedback effects between the price of the asset and the price of the derivative, which results in increased volatility.

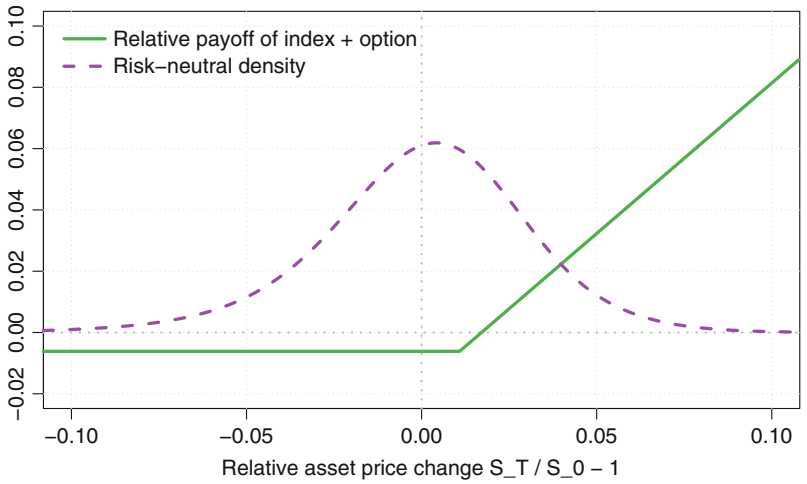


Fig. 2 Relative payoff r_h of an option-hedged portfolio example at maturity T and risk-neutral density of the underlying estimated at $t < T$ (scaled for visualization). The minimal loss of the hedged portfolio is $\min(r_h) = -0.6\%$

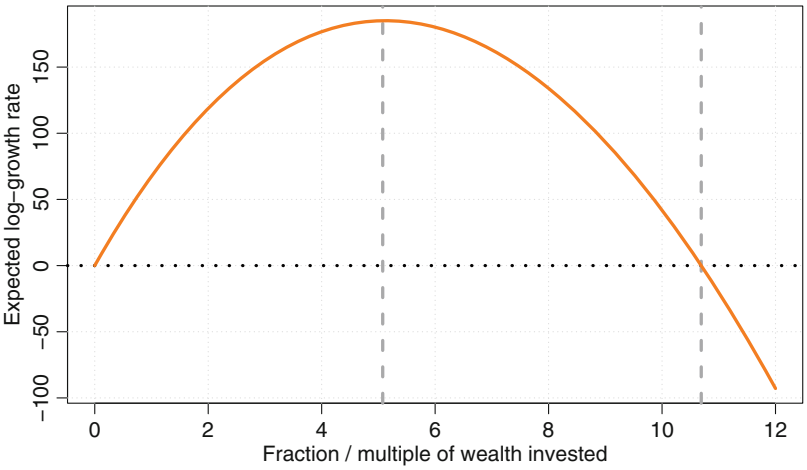


Fig. 3 Option-implied expected logarithmic growth rate of option-hedged portfolio example. The right zero crossing equals the Foster-Hart riskiness $FH(r_h) = 10.7$, the maximum growth rate the Kelly criterion $\alpha_K(r_h) = 5.1$

4 Data and Methods

In this section we discuss our data and the statistical methods employed in the empirical analysis.

4.1 Data

We obtain end-of-day bids, asks and open interest for standard European SPX call and put options on the S&P 500 stock market index for the period January 1st, 2003, to October 23rd, 2013, from Stricknet.[3] Throughout this decade the average daily market volume of SPX options grew from 150 to 890 K contracts and the open interest from 3840 to 11,883 K, respectively. In this study, we focus on monthly options, which are AM-settled and expire on the third Friday of a month. In addition, we use daily values for the S&P 500, its dividend yield, interest rates of 3-Month Treasury bills as a proxy of the risk-free rate, the (Chicago Board Options Exchange, 2009) Volatility Index (VIX) and the LIBOR from the Thomson Reuters Datastream.

4.2 Risk-Neutral Densities

Our first step is to extract risk-neutral densities from the option data as a market view on the probability distribution of the underlying gamble (which for our real-world finance application is of course unknown). There is a large literature on estimating risk-neutral probability distributions (Jackwerth, 2004). Here, we use our own method from Leiss et al. (2015), Leiss and Nax (2015) who generalize Figlewski (2010) for a modern, model-free method. We start with the fundamental theorem of asset pricing that states that in a complete market, the current price of an asset may be determined as the discounted expected value of the future payoff under the unique risk-neutral measure (e.g., Delbaen and Schachermayer, 1994). In particular, the price C_t of a standard European call option at time t with exercise price K and maturity T on a stock with price S is given as

$$C_t(K) = \mathrm{e}^{-r_f(T-t)}\,\mathbb{E}_t^{\mathbb{Q}}\left[\max(S_T - K, 0)\right] = \mathrm{e}^{-r_f(T-t)}\int_K^{\infty}(S_T - K) f_t(S_T)\mathrm{d}S_T, \tag{10}$$

where $\mathbb{Q}$ and f_t are the risk-neutral measure and the corresponding risk-neutral probability density, respectively. Since option prices C_t, the risk-free rate, r_f, and time to maturity, $T-t$ are observable, we can invert the pricing Eq. (10) to obtain an estimate for the risk-neutral density f_t. In practice, this involves numerical evaluation of derivatives (Breeden and Litzenberger, 1978) and fitting in implied volatility space (Shimko et al., 1993). Outside of the range of observable strike prices we fit tails of the family of generalized extreme value distributions, which are well-suited for the modeling extreme events (Embrechts et al., 1997). We refer the more interested reader to Figlewski (2010); Leiss et al. (2015); Leiss and Nax (2015) for details of the method.

[3]The data is available for purchase at http://www.stricknet.com/. More information on the SPX option contract specifications can be found at http://www.cboe.com/SPX.

4.3 Leverage Ratios

We will use the option-implied Foster-Hart riskiness of levered investments $FH^{\mathbb{Q}}(r_h)$ with r_h defined in (7) to estimate the prevailing leverage ratio. We compute $FH^{\mathbb{Q}}(r_h)$ for each business day and each put option available on that day. Be $\hat{P}_0$ the premium and $\hat{K}$ the exercise price with maximum $FH^{\mathbb{Q}}(r_h)$ on that business day. We report leverage ratios $FH^{\mathbb{Q}}(r_h(\hat{P}_0, \hat{K}))$ and, as a comparison, also the Kelly criterion $\alpha^{\mathbb{Q}}(r_h(\hat{P}_0, \hat{K}))$ as that quantity that numerically maximizes the option-implied logarithmic growth rate. Finally, we compute the future return $r_h(\hat{P}_0, \hat{K})$ with the realized value S_T of the underlying index at maturity.

4.4 Return Downturn Regression

We will assess the predictive power of risk measures with respect to extreme losses in the form of logistic regressions. For this, we define a binary downturn variable Δr_t^{ρ} that equals 1 in the case of an extreme event, and 0 otherwise:

$$\Delta r_t^{\rho} = \begin{cases} 1, & \text{if } r_{t\to T} < \rho, \\ 0, & \text{if } r_{t\to T} \geq \rho, \end{cases} \tag{11}$$

where ρ is a quantile describing the 5%, 10%, or 20% worst return. We note that $r_{t\to T}$ is the future *realized* return from time t to the maturity of the option T, and corresponds to the capital gain of a non-levered r_s (6) or levered portfolio r_h (7). In this sense our analysis allows inference about the predictive power of risk measures. We will regress downturns on individual risk measures R

$$\Delta r_t^{\rho} = a_{0,t} + a_{R,t}\, R_t + \varepsilon_t, \tag{12}$$

and on sets of risk measures $\mathscr{R}$:

$$\Delta r_t^{\rho} = a_{0,t} + \sum_{R\in\mathscr{R}} a_{R,t} R_t + \varepsilon_t. \tag{13}$$

Specifically, we will include the option-implied Foster-Hart riskiness $FH^{\mathbb{Q}}(r_h)$ and 5% Value at Risk of levered portfolios $VaR^{\mathbb{Q}}(r_h)$.[4] Leiss and Nax (2015) performed rigorous variable selection using the least absolute shrinkage and selection operator and found three further risk measures to be indicative (Tibshirani, 1996): (1) option-implied 5% expected shortfall of non-levered portfolios $ES^{\mathbb{Q}}(r_s)$, (2) the Chicago Board Options Exchange (2009) Volatility Index (VIX), and (3) the difference between the 3-month LIBOR and 3-month T-Bill rates (TED), a measure of credit risk. We will consider those indicators as well.

[4] Our results are robust with respect to choosing a different VaR level.

Over successive business days the downturns (11) focus on the same maturity T, as option exercise dates are standardized. This may induce autocorrelation in the dependent variable, which we correct for by using the heteroskedasticity and autocorrelation consistent covariance matrix estimators by Newey and West (1987, 1994).

5 Empirical Results

Having established the leveraged Foster-Hart riskiness and methods used, we now study empirical applications. First, we discuss the time dynamics of the option-implied leverage ratios around the Global Financial Crisis of 2008. Next, we analyze the predictive power of various risk measures with respect to extreme losses of levered and non-levered portfolios.

5.1 Option-Implied Leverage Around the Global Financial Crisis

Geanakoplos (2010) reports dramatically increased leverage from 1999 to 2006. In 2006, a bank could borrow as much as 98.4% of the purchase price of a AAA-rated mortgage-backed security, which corresponds to an average ratio of about 60 to 1. However, these numbers should not be directly compared to our findings, as the leverage ratios are defined differently. We assess leverage in time periods before and after the onset of the Global Financial Crisis, which Leiss et al. (2015) identified as June 22, 2007. Table 1 summarizes the option-implied Foster-Hart riskiness for non-levered $FH(r_s)$ and levered investments $FH(r_h)$. Prior to the Global Financial Crisis of 2008 the non-levered $FH(r_s)$ on average recommends investments of about 78% of one's wealth. During and after the crisis this value drops to about half its previous level.

In terms of FH-recommended leverage, we find an average leverage ratio of 105 in the pre-crisis regime, albeit with a fairly large confidence interval of ± 40 (see Fig. 4). During and after the crash it shrinks drastically to about 3.4. Geanakoplos

Table 1 Average levels of option-implied Foster-Hart riskiness, levered Foster-Hart riskiness, and Kelly criterion with 95% confidence intervals prior and after the onset of the Global Financial Crisis identified as 22 June 2007

Pre-crisis			Crisis and post-crisis
Non-levered Foster-Hart riskiness	$FH^{\mathbb{Q}}(r_s)$	0.78 ± 0.03	0.40 ± 0.02
Levered Foster-Hart riskiness	$FH^{\mathbb{Q}}(r_h)$	105 ± 40	3.4 ± 0.8
Levered Kelly criterion	$\alpha_K^{\mathbb{Q}}(r_h)$	41.00 ± 0.03	1.57 ± 0.02

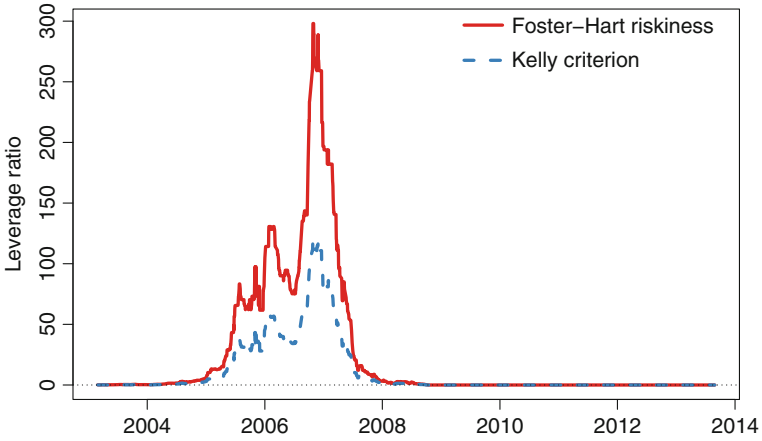


Fig. 4 Leverage according to option-implied Foster-Hart riskiness and Kelly criterion of leveraged gambles. Leverage ratios rise to drastically high values during the boom in mortgage-backed securities prior to 2008

(2010) explains the extraordinarily high leverage ratios during the pre-crisis years by financial innovation, namely the extensive use and abuse of credit default swaps (CDS). CDS are a vehicle for speculators to leverage their beliefs. Their standardization for mortgages led to enormous CDS trading prior at the peak of the housing bubble. Another reason for pronounced leverage before the crisis is the existence of two mutually reinforcing leverage cycles in mortgage-backed securities and housing (Geanakoplos, 2010). The option-implied Kelly criterion of hedged portfolios $\alpha_K^{\mathbb{Q}}(r_h)$recommends a leverage of 41 pre-crisis and 1.57 afterwards, with respective small confidence intervals of 0.03 and 0.02.

5.2 *Option-Implied Leveraged Foster-Hart Riskiness and Downturns*

We now assess the predictive power of various risk measures with respect to extreme future losses. Leiss and Nax (2015) empirically demonstrated that both Foster-Hart riskiness $FH(r_s)$ and the TED spread predict future downturns of non-hedged portfolios. Here, we will be specifically interested in the situation when the non-levered $FH(r_s)$ is stuck at the hard bound of 1 and therefore may only yield limited information. Thus, we subset our data to the 740 business days in our time period where $FH(r_s) = 1$.

Table 2 summarizes regression results for the 5%, 10%, 20% worst losses. We find that the option-implied Foster-Hart riskiness of levered portfolios helps predicting future downturns for very extreme events (at the 5% quantile and below). In the case of the 10% most negative performances, the option-implied value at risk

Table 2 Regressions of option-hedged portfolio downturns on various risk measures over 740 observations

Regression of the worst 5% downturns on risk measures (37 events)							
(Intercept)	−1.565***	−4.483***	−3.975***	−2.825***	−6.454***	−4.156***	−4.763***
	(0.273)	(0.312)	(0.342)	(0.416)	(0.735)	(1.046)	(1.123)
$-FH^{\mathbb{Q}}(r_h)$	0.263***					0.218**	0.138*
	(0.073)					(0.074)	(0.064)
$VaR^{\mathbb{Q}}(r_h)$		61.166***					20.326
		(8.661)					(13.197)
$ES^{\mathbb{Q}}(r_s)$			0.329***			0.129	0.164
			(0.081)			(0.131)	(0.156)
TED				−0.183		−0.452	−0.405
				(0.548)		(0.525)	(0.519)
VIX					0.194***	0.119	0.098
					(0.035)	(0.073)	(0.070)
Regression of the worst 10% downturns on risk measures (74 events)							
(Intercept)	−1.014***	−3.715***	−3.184***	−2.238***	−5.310***	−2.668***	−3.661***
	(0.276)	(0.258)	(0.335)	(0.361)	(0.635)	(0.792)	(0.860)
$-FH^{\mathbb{Q}}(r_h)$	0.128					0.103	0.046
	(0.066)					(0.070)	(0.034)
$VaR^{\mathbb{Q}}(r_h)$		64.658***					37.721***
		(7.962)					(11.266)
$ES^{\mathbb{Q}}(r_s)$			0.351***			0.201	0.272
			(0.105)			(0.138)	(0.166)
TED				0.110		−0.225	0.024
				(0.508)		(0.440)	(0.414)
VIX					0.178***	0.057	0.017
					(0.034)	(0.057)	(0.061)
Regression of the worst 20% downturns on risk measures (148 events)							
(Intercept)	−0.439*	−2.328***	−2.287***	−1.831***	−4.889***	−2.605**	−2.948***
	(0.214)	(0.240)	(0.292)	(0.337)	(0.738)	(0.831)	(0.823)
$-FH^{\mathbb{Q}}(r_h)$	0.048**					0.040*	0.030
	(0.018)					(0.018)	(0.017)
$VaR^{\mathbb{Q}}(r_h)$		49.538***					16.305
		(8.005)					(10.774)
$ES^{\mathbb{Q}}(r_s)$			0.358**			0.083	0.094
			(0.113)			(0.099)	(0.105)
TED				0.778		0.084	0.229
				(0.400)		(0.424)	(0.415)
VIX					0.207***	0.101	0.087
					(0.044)	(0.054)	(0.053)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The dependent variable reflects if the realized ahead-return of an option-hedged portfolio belongs to the set of the worst 5% (top panel), 10% (middle), or 20% (bottom) downturns in that period. The risk measures involve the option-implied Foster-Hart riskiness $FH^{\mathbb{Q}}(r_h)$ and value at risk $VaR^{\mathbb{Q}}(r_h)$ of the hedged portfolio, the option-implied expected shortfall $ES^{\mathbb{Q}}(r_s)$ of the non-hedged portfolio, as well as the industry measures TED spread (credit risk) and the volatility index VIX

Table 3 Regressions of stock market downturns on various risk measures over 740 observations

Regression of the worst 5% index downturns on risk measures							
(Intercept)	-2.369^{***}	-3.460^{***}	-3.999^{***}	-3.820^{***}	-6.666^{***}	-4.450^{***}	-4.492^{**}
	(0.296)	(0.423)	(0.410)	(0.367)	(0.752)	(1.318)	(1.429)
$FH^{\mathbb{Q}}(r_h)$	0.022^{**}					0.016^{***}	0.015^{***}
	(0.008)					(0.005)	(0.004)
$VaR^{\mathbb{Q}}(r_h)$		26.867^{**}					2.595
		(10.186)					(11.657)
$ES^{\mathbb{Q}}(r_s)$			0.336^{***}			0.281^{*}	0.282^{*}
			(0.079)			(0.134)	(0.138)
TED				1.315^{**}		0.862	0.886
				(0.500)		(0.624)	(0.571)
VIX					0.204^{***}	0.022	0.020
					(0.038)	(0.102)	(0.101)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The dependent variable reflects if the realized ahead-return of the S&P 500 stock market index belongs to the set of the worst 5% downturns in that period. The risk measures involve the option-implied Foster-Hart riskiness $FH^{\mathbb{Q}}(r_h)$ and value at risk $VaR^{\mathbb{Q}}(r_h)$ of the hedged portfolio, the option-implied expected shortfall $ES^{\mathbb{Q}}(r_s)$ of the non-hedged portfolio, as well as the industry measures TED spread (credit risk) and the volatility index VIX

of levered portfolios shows to be a significant predictor. Including even less extreme events, we find that while individually risk measures remain predictively successful, they lose significance in a joint regression.

Finally, we study if risk measures inferred from levered portfolios contain information about the future performance of non-levered investments. Table 3 summarizes our findings. The Foster-Hart riskiness estimated for hedged returns significantly explains future drops of simple returns both individually and in a joint regression. The same is true for the expected shortfall of non-levered investments as already documented in Leiss and Nax (2015).

6 Conclusion

In this paper we discussed a theoretical extension of the Foster-Hart measure of riskiness to study leverage. Option hedging prevents the value of portfolios from vanishing completely (provided the seller of the option does not default). In turn, this "frees" the Foster-Hart riskiness measure to values larger than 1, i.e. allows for leverage. Based on options data, we applied this new way of estimating prevailing leverage ratios to the decade 2003–2013 around the Global Financial Crisis. We found (1) a strong procyclicality of leverage during the bubble prior to the crash and (2) predictive power of risk measures computed for levered portfolios with respect to extreme losses.

Acknowledgements Leiss acknowledges support from the ETH Risk Center and through SNF grant *The Anatomy of Systemic Financial Risk*, Nax from the European Commission through the ERC Advanced Investigator Grant *Momentum* (Grant No. 324247).

References

Adrian, T., Shin, H.S.: Liquidity and leverage. J. Financ. Intermed. **19**(3), 418–437 (2010)
Adrian, T., Shin, H.S.: Procyclical leverage and value-at-risk. Rev. Financ. Stud. **27**(2), 373–403 (2014)
Anand, A., Li, T., Kurosaki, T., Kim, Y.S.: Foster–Hart optimal portfolios. J. Bank. Financ. **68**, 117–130 (2016)
Aumann, R.J., Serrano, R.: An economic index of riskiness. J. Polit. Econ. **116**(5), 810–836 (2008)
Basel Committee on Banking Supervision: Basel III: a global regulatory framework for more resilient banks and banking systems. Technical report, Bank for International Settlements (2010)
Black, F., Scholes, M.: The pricing of options and corporate liabilities. J. Polit. Econ. **81**(3), 637–654 (1973)
Breeden, D.T., Litzenberger, R.H.: Prices of state-contingent claims implicit in option prices. J. Bus. **51**(4), 621–651 (1978)
Chicago Board Options Exchange: The CBOE volatility index – VIX. Technical report, White Paper (2009)
Delbaen, F., Schachermayer, W.: A general version of the fundamental theorem of asset pricing. Math. Ann. **300**(1), 463–520 (1994)
Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events: For Insurance and Finance, vol. 33. Springer, Berlin (1997)
Embrechts, P., Frey, R., McNeil, A.: Quantitative Risk Management, vol. 10. Princeton Series in Finance, Princeton (2005)
Figlewski, S.: Estimating the implied risk neutral density. In: Bollerslev, T., Russell, J., Watson, M. (eds.) Volatility and Time Series Econometrics. Oxford University Press, Oxford (2010)
Foster, D.P., Hart, S.: An operational measure of riskiness. J. Polit. Econ. **117**(5), 785–814 (2009)
Geanakoplos, J.: The leverage cycle. In: NBER Macroeconomics Annual 2009, vol. 24, pp. 1–65. University of Chicago Press, Chicago (2010)
Gorton, G., Metrick, A.: Securitized banking and the run on repo. J. Financ. Econ. **104**(3), 425–451 (2012)
Grossman, S.J., Vila, J.-L.: Optimal dynamic trading with leverage constraints. J. Financ. Quant. Anal. **27**(02), 151–168 (1992)
Hadar, J., Russell, W.R.: Rules for ordering uncertain prospects. Am. Econ. Rev. **59**(1), 25–34 (1969)
Hanoch, G., Levy, H.: The efficiency analysis of choices involving risk. Rev. Econ. Stud. **36**(3), 335–346 (1969)
Hildebrand, P.M.: Is Basel II Enough? The Benefits of a Leverage Ratio. Philipp M. Hildebrand, Vice-Chairman of the Governing Board Swiss National Bank, in a Financial Markets Group Lecture at the London School of Economics on December 15, 2008. http://www.ub.unibas.ch/digi/a125/sachdok/2011/BAU_1_5654573.pdf (2008)
Jackwerth, J.C.: Option-Implied Risk-Neutral Distributions and Risk Aversion. Research Foundation of AIMR Charlotteville (2004)
Kadan, O., Liu, F.: Performance evaluation with high moments and disaster risk. J. Financ. Econ. **113**(1), 131–155 (2014)
Kelly, J.L.: A new interpretation of information rate. IRE Trans. Inf. Theory **2**(3), 185–189 (1956)
Leiss, M., Nax, H.H.: Option-implied objective measures of market risk. Social Science Research Network Working Paper Series, 2690476, Quantitative Economics (2015, submitted)

Leiss, M., Nax, H.H., Sornette, D.: Super-exponential growth expectations and the global financial crisis. J. Econ. Dyn. Control **55**, 1–13 (2015)
Newey, W.K., West, K.D.: A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica **55**(3), 703–708 (1987)
Newey, W.K., West, K.D.: Automatic lag selection in covariance matrix estimation. Rev. Econ. Stud. **61**(4), 631–653 (1994)
Riedel, F., Hellmann, T.: The Foster-Hart measure of riskiness for general gambles. Theor. Econ. **10**(1), 1–9 (2015)
Rothschild, M., Stiglitz, J.E.: Increasing risk: I. A definition. J. Econ. Theory **2**(3), 225–243 (1970)
Schularick, M., Taylor, A.M.: Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870–2008. Am. Econ. Rev. **102**(2), 1029–1061 (2012)
Sharpe, W.F.: The sharpe ratio. J. Portf. Manag. **21**(1), 49–58 (1994)
Shimko, D.C., Tejima, N., Van Deventer, D.R.: The pricing of risky debt when interest rates are stochastic. J. Fixed Income **3**(2), 58–65 (1993)
Sircar, R.K., Papanicolaou, G.: General Black-Scholes models accounting for increased market volatility from hedging strategies. Appl. Math. Finance **5**(1), 45–82 (1998)
Thurner, S., Farmer, J.D., Geanakoplos, J.: Leverage causes fat tails and clustered volatility. Quant. Finan. **12**(5), 695–707 (2012)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. **58**(1), 267–288 (1996)
Von der Becke, S., Sornette, D.: Toward a unified framework of credit creation. Technical Report 14-07, Swiss Finance Institute Research Paper (2014)
Whitworth, W.: Choice and Chance. Deighton, Bell and Co, Cambridge (1870)

The Sustainable Black-Scholes Equations

Yannick Armenti, Stéphane Crépey, and Chao Zhou

Abstract In incomplete markets, a basic Black-Scholes perspective has to be complemented by the valuation of market imperfections. Otherwise this results in Black-Scholes Ponzi schemes, such as the ones at the core of the last global financial crisis, where always more derivatives need to be issued for remunerating the capital attracted by the already opened positions. In this paper we consider the sustainable Black-Scholes equations that arise for a portfolio of options if one adds to their trade additive Black-Scholes price, on top of a nonlinear funding cost, the cost of remunerating at a hurdle rate the residual risk left by imperfect hedging. We assess the impact of model uncertainty in this setup.

Keywords Market incompleteness • Cost of capital (KVA) • Cost of funding (FVA) • Model risk • Volatility uncertainty • Optimal martingale transport

1 Introduction

In incomplete markets, a basic Black-Scholes perspective has to be complemented by the valuation of market imperfections. Otherwise this results in Black-Scholes Ponzi schemes, such as the ones at the core of the last global financial crisis, where always more derivatives need to be issued for remunerating the capital attracted by the already opened positions. In this paper we consider the sustainable Black-Scholes equations that arise for a portfolio of options if one adds to their trade additive Black-Scholes price, on top of a nonlinear funding cost, the cost of remunerating at a hurdle rate the residual risk left by imperfect hedging. We assess the impact of model uncertainty in this setup.

Section 2 revisits the pricing of a book of options accounting for cost of capital and cost of funding, which are material in incomplete markets. Section 3 specializes

Y. Armenti • S. Crépey (✉)
University of Evry Val d'Essonne, Evry, France
e-mail: stephane.crepey@univ-evry.fr

C. Zhou
Department of Mathematics, National University of Singapore, Singapore, Singapore
e-mail: matzc@nus.edu.sg.

J.A. Londoño et al. (eds.), *Actuarial Sciences and Quantitative Finance*, Springer
Proceedings in Mathematics & Statistics 214, https://doi.org/10.1007/978-3-319-66536-8_8

the pricing equations to a Markovian Black–Scholes setup. Section 4 assesses the impact of model risk in a UVM (uncertain volatility model) setup. Section 5 refines the model risk add-ons by accounting for calibrability constraints.

We consider a portfolio of options made of ω_i vanilla call options of maturity T_i and strike K_i on a stock S, with $0 < T_1 < \ldots < T_n = T$. Note that, if a corporate holds a bank payable, it typically has an appetite to close it, receive cash, and restructure the hedge otherwise with a par contract (the bank would agree to close the deal as a market maker, charging fees for the new trade). Because of this natural selection, a bank is mostly in the receivables (i.e. "$\omega_i \geq 0$") in its derivative business with corporates.

We write $x^{\pm} = \max(\pm x, 0)$.

2 Cost of Capital and Cost of Funding

2.1 Cost of Capital

In presence of hedging imperfections resulting in a nonvanishing loss (and profit) process ϱ of the bank, a conditional risk measure $\mathrm{EC} = \mathrm{EC}_t(\varrho)$ must be dynamically computed and reserved by the bank as economic capital.

It is established in Albanese et al. (2016, Sect. 5) that the capital valuation adjustment (KVA) needed by the bank in order to remunerate its shareholders for their capital at risk at some average hurdle rate h (e.g. 10%) at any point in time in the future is:

$$\mathrm{KVA} = \mathrm{KVA}_t(\varrho) = h\mathbb{E}_t \int_t^T e^{-(r+h)(s-t)} \mathrm{EC}_s(\varrho) ds, \tag{1}$$

where $\mathbb{E}_t$ stands for the conditional expectation with respect to some probability measure $\mathbb{Q}$ and model filtration.

In principle, the probability measure used in capital and cost of capital calculations should be the historical probability measure. But, in the present context of optimization of a portfolio of derivatives, the historical probability measure is hard to estimate in a relevant way, especially for long maturities. As a consequence, we do all our price and risk computations under a risk-neutral measure $\mathbb{Q}$ calibrated to the market (or a family of pricing measures, in the context of model uncertainty later below), assuming no arbitrage.

2.2 Cost of Funding

Let r_t denote a risk-free OIS short-term interest rate and $\beta_t = e^{-\int_0^t r_s ds}$ be the corresponding risk-neutral discount factor. We assume that the bank can invest at the risk-free rate r but can only obtain unsecured funding at a shifted rate $r + \lambda > r$.

This entails funding costs over OIS and a related funding valuation adjustment (FVA) for the bank. Given our focus on capital and funding in this paper, we ignore counterparty risk for simplicity, so that λ is interpreted as a pure funding liquidity basis. In order to exclude arbitrages in the primary market of hedging instruments, we assume that the vector gain process $\mathcal{M}$ of unit positions held in the hedging assets is a risk-neutral martingale. The bank "marks to the model" its derivative portfolio, assumed bought from the client at time 0, by means of an FVA-deducted value process Θ. The bank may also set up a (possibly imperfect) hedge $(-\eta)$ in the hedging assets, for some predictable row-vector process η of the same dimension as $\mathcal{M}$. We assume that the depreciation of Θ, the funding expenditures and the loss $\eta d\mathcal{M}$ on the hedge, minus the option payoffs as they mature, are instantaneously realized into the loss(-and-profit) process ϱ of the bank. In particular, at any time t, the amount on the funding account of the bank is Θ_t. Moreover, we assume that the economic capital can be used by the trader for her funding purposes provided she pays to the shareholders the OIS rate on EC that they would make otherwise by depositing it (assuming it all cash for simplicity).

Note that the value process Θ of the trade already includes the FVA as a deduction, but ignores the KVA, which is considered as a risk adjustment computed in a second step (in other words, we assume that the trader's account and the KVA account are kept separate from each other). Rephrasing in mathematical terms the above description, the loss equation of the trader is written, for $t \in (0, T]$, as (starting from $\varrho_0 = y$, the accrued loss of the portfolio):

$$
\begin{aligned}
d\varrho_t = & \underbrace{-\sum_i \omega_i (S_{T_i} - K_i)^+ \delta_{T_i}(dt)}_{\text{call payoffs}} \\
& + \underbrace{r_t \mathrm{EC}_t(\varrho) dt}_{\substack{\text{Payment of internal lending of the EC funding source at OIS rate}}} \\
& + \underbrace{\Big((r_t + \lambda_t)\big(\Theta_t - \mathrm{EC}_t(\varrho)\big)^+ - r_t\big(\Theta_t - \mathrm{EC}_t(\varrho)\big)^- \Big) dt}_{\text{portfolio funding costs/benefits}} \qquad (2) \\
& + \underbrace{(-d\Theta_t)}_{\substack{\text{depreciation of } \Theta}} + \underbrace{\eta_t d\mathcal{M}_t}_{\text{loss on the hedge}} \\
= & -d\Theta_t - \sum_i \omega_i (S_{T_i} - K_i)^+ \delta_{T_i}(dt) + \Big(\lambda_t \big(\Theta_t - \mathrm{EC}_t(\varrho)\big)^+ + r_t \Theta_t \Big) dt + \eta_t d\mathcal{M}_t .
\end{aligned}
$$

Hence, a no-arbitrage condition that the loss process ϱ of the bank should follow a risk-neutral martingale (assuming integrability) and the terminal condition $\Theta_T = 0$ lead to the following FVA-deducted risk-neutral valuation BSDE:

$$
\Theta_t = \underbrace{\mathbb{E}_t\Big[\sum_{t<T_i} \beta_t^{-1} \beta_{T_i} \omega_i (S_{T_i} - K_i)^+ \Big]}_{\Theta_t^0} - \underbrace{\mathbb{E}_t\Big[\int_t^T \beta_t^{-1} \beta_s \lambda_s \big(\Theta_s - \mathrm{EC}_s(\varrho)\big)^+ ds \Big]}_{\mathrm{FVA}_t}, \ t \in [0, T] \qquad (3)
$$

(since we consider a portfolio of options with several maturities, we treat option pay-offs as cash-flows at their maturity times rather than a terminal condition in the equations, in particular $\Theta_T = 0$).

The funding source provided by economic capital creates a feedback loop from EC into FVA, which makes the FVA smaller.

Note that, in the usual case of a risk measure EC only affected by the time fluctuations of ϱ, the Eqs. (3) and in turn (1) are independent of the accrued loss y, which eventually does not affect Θ nor the KVA.

If $\lambda = 0$, then, whatever the hedge η, Θ reduces to Θ^0, which corresponds to the usual trade additive (linear) no-arbitrage pricing formula for a portfolio of options, with zero FVA, but with a KVA given by (1), depending on the hedge η.

If $\lambda \neq 0$, we introduce the following backward SDE:

$$\Theta_t^\star = \mathbb{E}_t\Big[\sum_{t<T_i} \beta_t^{-1}\beta_{T_i}\omega_i(S_{T_i} - K_i)^+ - \int_t^T \beta_t^{-1}\beta_s\lambda_s(\Theta_s^\star)^+ ds\Big],\ t \in [0,T]\,. \tag{4}$$

This is a monotone driver backward SDE, admitting as such a unique square integrable solution $\Theta^\star$ (see, e.g., Kruse and Popier (2016, Sect. 4)), provided λ is bounded from below and Θ^0 is square integrable. If there exists a replicating hedge η, i.e. $\eta = \eta^\star$ such that the corresponding ϱ is constant in (2), i.e. $\eta_t^\star d\mathcal{M}_t$ coincides with the martingale part of $\Theta^\star$, then the resulting ϱ, EC and KVA vanish (since we assumed $\mathrm{EC}(0) = 0$) and the ensuing FVA-deducted value process is given by $\Theta^\star$.

Example 2.1 (Single option positions) If $n = 1$ and $\omega_1 = 1$ (one long call position), then, by application of the comparison theorem for BSDEs with a monotonic generator (see Kruse and Popier (2016, Sect. 4)), we have $\Theta^\star \geq 0$, hence

$$\Theta_t^\star = \mathbb{E}_t\big[\widetilde{\beta}_t^{-1}\widetilde{\beta}_{T_1}(S_{T_1} - K_1)^+\big], \tag{5}$$

where $\widetilde{\beta}_t = e^{-\int_0^t (r+\lambda_s)ds}$. With respect to $\Theta^{(0)}$, the value $\Theta^\star$ corresponds to an FVA rebate on the buying price by the bank (since we assumed a positive liquidity basis λ).

If $n = \omega_1 = -1$ (one short call position), then we deduce likewise that $\Theta^\star \leq 0$, hence $\Theta^\star = \Theta^{(0)}$.

But, apart from the above special cases where $\lambda = 0$ or $\eta = \eta^\star$, the BSDE (3) for Θ is nonstandard due to the term $\mathrm{EC} = \mathrm{EC}_t(\varrho)$ in the FVA.

3 Markovian Black-Scholes Setup

In this section we assume a constant risk-free rate r and a Black-Scholes stock S with volatility σ and constant dividend yield q. The risk-neutral martingale $\mathcal{M}$ is then taken as the gain process of a continuously rolled unit position on the stock S,

assumed funded at the risk-free rate via a repo market, i.e. $d\mathcal{M}_t = dS_t - (r-q)S_t dt$. We denote by $\mathcal{A}_S^{bs} = (r-q)S\partial_S + \frac{1}{2}\sigma^2 S^2 \partial_{S^2}^2$ the corresponding risk-neutral Black-Scholes generator.

Doing our modeling exercise in the context of the Black-Scholes model, where perfect replication, hence no KVA, is possible, may seem rather artificial. However, doing all the computations in a stylized Black-Scholes setup with a single risk factor S yields useful practical insights. In addition, this conveys the message that, in real-life incomplete markets, a basic Black-Scholes perspective has to be complemented by the valuation of market imperfections, otherwise this unavoidably results in Black-Scholes Ponzi schemes, such as the ones that have been involved in the global financial crisis, where always more derivatives are issued to remunerate the capital required by the already opened positions (if priced and risk-managed in a basic Black-Scholes way ignoring the cost of capital).

In the Black-Scholes setup and assuming a stylized Markovian specification

$$\mathrm{EC}_t(\varrho) = f\sqrt{\frac{d\langle \varrho \rangle}{dt}} \tag{6}$$

(the stylized VaR which is proportional to the instantaneous volatility of the loss process ϱ modulo a suitable "quantile level" f) as well as $\lambda = \lambda(t, S_t)$, $\eta_t = \eta(t, S_t)$, then the above FVA and KVA equations can be reduced to the "sustainable Black-Scholes PDEs" (12), as follows (resulting in an FVA- and KVA-deducted price that would be sustainable for the bank even in the limit case of a portfolio held on a run-off basis, with no new trades ever entered in the future).

First, observe that given a tentative FVA-deducted price process of the form $\Theta_t = u(t, S_t)$ for some to-be-determined function $u = u(t, S)$, we have, assuming (6):

$$\sqrt{\frac{d\langle \varrho \rangle}{dt}} = \sigma S_t \big| \partial_S u(t, S_t) - \eta(t, S_t) \big|. \tag{7}$$

Accordingly, let the function u be defined by $u_i(t, S)$ on each strip $(T_{i-1}, T_i] \times (0, \infty)$, where $(u_i)_{1 \le i \le n}$ is the unique sequence of viscosity solutions, which can then be shown to be classical solutions, to the following PDE cascade, for i decreasing from n to 1 (closing the system by setting $u_{n+1} = 0$ and $T_0 = 0$):

$$\begin{cases} u_i(T_i, S) = u_{i+1}(T_i, S) + \omega_i (S - K_i)^+ \text{ on } (0, \infty) \\ \partial_t u_i + \mathcal{A}_S^{bs} u_i - \lambda \big(u_i - f\sigma S |\partial_S u_i - \eta| \big)^+ - r u_i = 0 \text{ on } [T_{i-1}, T_i) \times (0, \infty). \end{cases} \tag{8}$$

Itô calculus shows that the process $\Theta = (u(t, S_t))_t$ solves the Markovian, monotonic driver (assuming λ bounded from below) BSDE

$$\begin{aligned} u(t, S_t) = \mathbb{E}_t \Big[& \sum_{t < T_i} \beta_t^{-1} \beta_{T_i} \omega_i (S_{T_i} - K_i)^+ \\ & - \int_t^T \beta_t^{-1} \beta_s \lambda_s \Big(u(s, S_s) - f\sigma S_s \big| \partial_S u(s, S_s) - \eta(s, S_s) \big| \Big)^+ ds \Big], \; t \in [0, T], \end{aligned} \tag{9}$$

which in view of (6)–(7) is precisely (3).

The ensuing FVA$=\Theta^{(0)}-\Theta$ and KVA processes are given as (cf. (3) and (1)):

$$\begin{aligned}\mathrm{FVA}_t(\varrho) &= \mathbb{E}_t\Big[\int_t^T e^{-r(s-t)}\lambda_s\Big(u(s,S_s)-f\sqrt{\frac{d\langle\varrho\rangle}{ds}}\Big)^+ ds\Big]\\ \mathrm{KVA}_t(\varrho) &= h\mathbb{E}_t\Big[\int_t^T e^{-(r+h)(s-t)} f\sqrt{\frac{d\langle\varrho\rangle}{ds}} ds\Big],\end{aligned} \tag{10}$$

where $\sqrt{\frac{d\langle\varrho\rangle}{dt}}$ is given by (7). We set $\eta=(1-\alpha)\partial_S u$, where α in $[0,100\%]$ is the mis-hedge parameter (noting that, for $\alpha=0$, the BSDE (9) reduces to the replication BSDE (4)), then the latter reduces to $\alpha\sigma S_t\big|\partial_S u(t,S_t)\big|$ and we have

$$\begin{aligned}\mathrm{FVA}_t(\varrho) &= \mathbb{E}_t\Big[\int_t^T e^{-r(s-t)}\lambda_s\big(u(s,S_s)-\alpha f\sigma S_s\big|\partial_S u(s,S_s)\big|\big)^+ ds\Big]\\ &= v(t,S_t)=u_{bs}(t,S_t)-u(t,S_t),\\ \mathrm{KVA}_t(\varrho) &= h\mathbb{E}_t\Big[\int_t^T e^{-(r+h)(s-t)}\alpha f\sigma S_s\big|\partial_S u(s,S_s)\big| ds\Big]=w(t,S_t),\end{aligned} \tag{11}$$

where u_{bs} is the trade additive Black-Scholes portfolio value and where the FVA and KVA pricing functions v and w satisfy

$$\begin{cases} v(T,S)=w(T,S)=0 \text{ on } (0,\infty)\\ \partial_t v+\mathcal{A}_S^{bs}v+\lambda\big(u_{bs}-v-\alpha f\sigma S|\Delta_{bs}-\partial_S v|\big)^+-rv=0 \text{ on } [0,T)\times(0,\infty)\\ \partial_t w+\mathcal{A}_S^{bs}w+\alpha hf\sigma S|\Delta_{bs}-\partial_S v|-(r+h)w=0 \text{ on } [0,T)\times(0,\infty),\end{cases} \tag{12}$$

in which $\Delta_{bs}=\partial_S u_{bs}$.

These "sustainable Black-Scholes PDEs" (12) allow computing an FVA and KVA deducted price

$$u-w=u_{bs}-v-w$$

that would be sustainable for the bank even in the limit case of a portfolio held on a run-off basis, with no new trades ever entered in the future.

4 With Volatility Uncertainty

An important and topical issue, referred to by the regulation as AVA (additional valuation adjustment), is the magnifying impact of model risk on the different XVA metrics.

In this section, we assess model risk from the angle of Avellaneda et al. (1995)'s uncertain volatility model (UVM). Namely, we only assume positive bounds $\underline{\sigma}$ and $\overline{\sigma}$ but we do not assume any specific dynamic on the stock volatility process σ. Therefore, there is a model uncertainty about it. That is, we only consider $d\mathcal{M}_t := \sigma_t S_t dW_t = dS_t - (r-q)S_t dt$, where $\sigma_t \in [\underline{\sigma}, \overline{\sigma}]$ for every t.

We call $\mathcal{C}$ the space of continuous paths on $\mathbb{R}_+$, C the canonical process on the space $\mathcal{C}$, $\mathbb{F} = (\mathcal{F}_t)_{0\le t\le T}$ the canonical filtration generated by C and $\mathcal{Q}$ the set of $\mathbb{F}$ local martingale probability measures for C. We recall from Soner et al. (2012) that, for any probability measure $\mathbb{Q} \in \mathcal{Q}$, the process C satisfies $dC_t = a_t^{1/2} dW_t^{\mathbb{Q}}$, for some $\mathbb{Q}$ Brownian motion $W^{\mathbb{Q}}$, where a_t is the Lebesgue density of the aggregated quadratic variation of C. In the following we restrict attention to the probability measures $\mathbb{Q}$ such that $a_t^{1/2} \in [\underline{\sigma}, \overline{\sigma}]$ holds $dt \times \mathbb{Q}$ almost surely, still denoting by $\mathcal{Q}$ the (restricted) set of measures, and we model $d\mathcal{M}_t = dS_t - (r-q)S_t dt$ as $S_t dC_t$.

Under each $\mathbb{Q}$, similarly to (2), the loss equation of the trader is written, for $t \in (0, T]$, as:

$$\begin{aligned} d\varrho_t^{\mathbb{Q}} = -d\Theta_t^{\mathbb{Q}} - \sum_i \omega_i (S_{T_i} - K_i)^+ \delta_{T_i}(dt) + \Big(\lambda_t \big(\Theta_t^{\mathbb{Q}} - \mathrm{EC}_t^{\mathbb{Q}}(\varrho^{\mathbb{Q}})\big)^+ \\ + r_t \Theta_t^{\mathbb{Q}}\Big) dt + \eta_t d\mathcal{M}_t \end{aligned} \tag{13}$$

where $\mathrm{EC}^{\mathbb{Q}}$ is some conditional risk measure under $\mathbb{Q}$. The ensuing equation for the $\mathbb{Q}$ FVA-deducted value $\Theta^{\mathbb{Q}}$ appears as

$$\begin{aligned} \Theta_t^{\mathbb{Q}} = \mathbb{E}_t^{\mathbb{Q}}\Big[\sum_{t<T_i} \beta_t^{-1} \beta_{T_i} \omega_i (S_{T_i} - K_i)^+ - \int_t^T \beta_t^{-1} \beta_s \lambda_s \big(\Theta_s^{\mathbb{Q}} - \mathrm{EC}_s^{\mathbb{Q}}(\varrho^{\mathbb{Q}})\big)^+ ds \Big], \\ t \in [0, T]\,. \end{aligned} \tag{14}$$

Under each $\mathbb{Q}$, the trader should value the derivative portfolio $\Theta_0^{\mathbb{Q}}$ at time 0 (or $\Theta_t^{\mathbb{Q}}$ at time t). However, due to the model uncertainty, the trader values it $\Theta_0 = \inf_{\mathbb{Q}\in\mathcal{Q}} \Theta_0^{\mathbb{Q}}$ (or at time t, $\Theta_t = \operatorname{ess\,inf}_{\mathbb{Q}\in\mathcal{Q}} \Theta_t^{\mathbb{Q}}$), which is a robust non-arbitrage price in the sense of Biagini et al. (2015).

At time t, $\mathrm{EC}_t^{\mathbb{Q}}(\varrho^{\mathbb{Q}})$ may depend on the whole future of the process $(\varrho_s^{\mathbb{Q}})$, $s \ge t$. This makes (14) a so-called anticipated BSDE under $\mathbb{Q}$ (ABSDE in the sense of Peng and Yang (2009)), with generator $\lambda_t\big(\Theta_t^{\mathbb{Q}} - \mathrm{EC}_t^{\mathbb{Q}}(\varrho^{\mathbb{Q}})\big)^+$, where $\Theta^{\mathbb{Q}}$ corresponds to the "Y-component" and $(d\varrho_s^{\mathbb{Q}} - \eta_s S_s dC_s)$ to the "Z-component" of the solution. However, in the Markovian setting of Sect. 3, $\mathrm{EC}_t^{\mathbb{Q}}(\varrho^{\mathbb{Q}})$ only depends on $(\varrho_t^{\mathbb{Q}})$ at time t, so that the ABSDE (14) reduces to a BSDE.

For taking model risk (i.e. the impact of several $\mathbb{Q}$) into consideration, we need the notion of second order BSDE. Wellposedness results regarding second order anticipated BSDEs are not yet available in the literature. Hence, we only give heuristic formulations in this regard. Namely, by analogy with the second order

BSDEs theory introduced by Soner et al. (2012), we should have the following representation, where $\mathbb{F}_+ = (\mathcal{F}_t^+)_{0\le t\le T}$ the right limit of $\mathbb{F}$, i.e. $\mathcal{F}_t^+ = \cap_{s>t}\mathcal{F}_s$ for all $t \in [0,T)$ and $\mathcal{F}_T^+ = \mathcal{F}_T$:

There exists a process ϱ such that, for each $\mathbb{Q} \in \mathcal{Q}$, ϱ is a $\mathbb{Q}$-local martingale and it $\mathbb{Q}-a.s.$ holds that

$$\begin{aligned} d\varrho_t = &-d\Theta_t - \sum_i \omega_i (S_{T_i} - K_i)^+ \delta_{T_i}(dt) \\ &+ \Big(\lambda_t \big(\Theta_t - \mathrm{EC}_t^{\mathbb{Q}}(\varrho)\big)^+ + r_t \Theta_t\Big) dt + \eta_t d\mathcal{M}_t + dA_t^{\mathbb{Q}}, \end{aligned} \tag{15}$$

where $\mathrm{EC}^{\mathbb{Q}}$ is some conditional risk measure and the family $\{A^{\mathbb{Q}}\}$ of non-decreasing processes satisfies the minimality condition

$$A_t^{\mathbb{Q}} = \operatorname*{ess\,inf}^{\mathbb{Q}}_{\mathbb{Q}' \in \mathcal{Q}(t,\mathbb{Q},\mathbb{F}_+)} \mathbb{E}^{\mathbb{Q}'} \left[A_T^{\mathbb{Q}'} \,\middle|\, \mathcal{F}_t^{\mathbb{Q}+} \right],\ 0 \le t \le T,\ \mathbb{Q}-a.s.,\ \forall \mathbb{Q} \in \mathcal{Q}, \tag{16}$$

where $\mathcal{Q}(t,\mathbb{Q},\mathbb{F}_+) := \left\{ \mathbb{Q}' \in \mathcal{Q},\ \mathbb{Q}' = \mathbb{Q} \text{ on } \mathcal{F}_t^+ \right\}$.

The corresponding equation for the FVA-deducted value Θ would appear as

$$\begin{aligned} \Theta_t = \operatorname*{ess\,inf}_{\mathbb{Q}' \in \mathcal{Q}(t,\mathbb{Q},\mathbb{F}_+)} \mathbb{E}_t^{\mathbb{Q}'} \Big[&\sum_{t<T_i} \beta_t^{-1} \beta_{T_i} \omega_i (S_{T_i} - K_i)^+ \\ &- \int_t^T \beta_t^{-1} \beta_s \lambda_s \big(\Theta_s - \mathrm{EC}_s^{\mathbb{Q}'}(\varrho)\big)^+ ds \Big],\ t \in [0,T],\ \mathbb{Q}-a.s. \end{aligned} \tag{17}$$

4.1 Equations in the Markovian Setting

By contrast, in the Markovian setting of Sect. 3 with VaR-like specification of Economic Capital, we can make rigorous statements. According to the second order BSDE theory introduced in Soner et al. (2012), the PDE (8) becomes:

$$\begin{cases} u_i(T_i, S) = u_{i+1}(T_i, S) + \omega_i (S - K_i)^+ \text{ on } (0,\infty) \\ \partial_t u_i + \inf\limits_{\sigma \in [\underline{\sigma}, \overline{\sigma}]} \left[\mathcal{A}_S^{bs} u_i - \lambda \big(u_i - f\sigma S |\partial_S u_i - \eta|\big)^+ \right] - r u_i = 0 \text{ on } [T_{i-1}, T_i) \times (0,\infty). \end{cases} \tag{18}$$

Let u be defined by $u_i(t,S)$ on each strip $(T_{i-1}, T_i] \times (0,\infty)$. The FVA can be defined as $\Theta^{\lambda=0} - \Theta$ and the ensuing KVA process is given as (cf. (3) and (1)):

$$\mathrm{KVA}_t(\varrho) = h \operatorname*{esssup}_{\mathbb{Q}' \in \mathcal{Q}(t,\mathbb{Q},\mathbb{F}_+)} \mathbb{E}_t^{\mathbb{Q}'} \Big[\int_t^T e^{-(r+h)(s-t)} f \sqrt{\frac{d\langle \varrho \rangle}{ds}} ds \Big],\ \mathbb{Q}\ a.s., \tag{19}$$

where $\sqrt{\frac{d\langle \varrho \rangle}{dt}} = a_t^{1/2} S_t \left| \partial_S u(t, S_t) - \eta(t, S_t) \right|$. In the case where $\eta = (1-\alpha)\partial_S u$, we obtain

$$\mathrm{KVA}_t(\varrho) = w(t, S_t),$$

where

$$\begin{cases} w(T, S) = 0 \text{ on } (0, \infty) \\ \partial_t w + \sup\limits_{\sigma \in [\underline{\sigma}, \overline{\sigma}]} \left[\mathcal{A}_S^{bs} w + \alpha h f \sigma S |\partial_S u| \right] - (r+h) w = 0 \text{ on } [0, T) \times (0, \infty), \end{cases} \tag{20}$$

in which (cf. (18))

$$\begin{cases} u_i(T_i, S) = u_{i+1}(T_i, S) + \omega_i (S - K_i)^+ \text{ on } (0, \infty) \\ \partial_t u_i + \inf\limits_{\sigma \in [\underline{\sigma}, \overline{\sigma}]} \left[\mathcal{A}_S^{bs} u_i - \lambda \left(u_i - \alpha f \sigma S |\partial_S u_i| \right)^+ \right] - r u_i = 0 \text{ on } [T_{i-1}, T_i) \times (0, \infty). \end{cases}$$

5 Optimal Transportation Approach

Since vanilla call options are liquidly traded, their time 0 price components

$$\mathbb{E}^{\mathbb{Q}} \left[\beta_{T_i} (S_{T_i} - K_i)^+ \right]$$

should not be seen as subject to model risk, but calibrated to the market. Hence, we need to refine our preliminary UVM assessment of model risk in order to account for these calibration constraints. For simplicity we consider a single call option (T, K) and we set $\lambda = 0$, focusing on KVA in this section. Hence, the system (18) reduces to a single PDE with $\lambda = 0$, with solution denoted by u.

(Tan and Touzi (2013)) consider the optimal transportation problem consisting of minimizing a cost among all continuous semimartingales with given initial and terminal distributions. They show an extension of the Kantorovich duality to this context and suggest a finite-difference scheme combined with the gradient projection algorithm to approximate the dual value. Their results can be applied to our setup as follows.

Let $\mu_0 = \delta_{S_0}$ denote the Dirac measure on the initial value of S_0 and let μ_T denote the marginal distribution of S_T, inferred by calibration to the market prices of all European call options with maturity T (assuming quotations available for all strikes). Let

$$\mathcal{Q}(\mu_0) = \{\mathbb{Q} \in \mathcal{Q} : \mathbb{Q} \circ S_0^{-1} = \mu_0\}, \ \mathcal{Q}(\mu_0, \mu_T) = \{\mathbb{Q} \in \mathcal{Q}(\mu_0) : \mathbb{Q} \circ S_T^{-1} = \mu_T\}.$$

From the Remark 2.3 in Tan and Touzi (2013), $\mathcal{Q}(\mu_0, \mu_T)$ is not empty in our setting.

The KVA with model uncertainty and terminal marginal constraint is defined as follows:

$$\mathrm{KVA}_0(\varrho) = h \sup_{\mathbb{Q}\in\mathcal{Q}(\mu_0,\mu_T)} \mathbb{E}^{\mathbb{Q}}\Big[\int_0^T e^{-(r+h)(s)} f \sqrt{\frac{d\langle\varrho\rangle}{ds}} ds\Big], \tag{21}$$

where ϱ represents the portfolio loss in this setting, that is, the loss and profit of the bank in a world with uncertain volatility subject to the law of S_T. However, it is not clear how to extrapolate the theory of Tan and Touzi (2013) to valuation at future time points when only the unconditional law of S_T is known. Hence for the sake of tractability we conservatively assume that ϱ in (21) is the UVM one and we only apply the constraint to the outer expectation in (21) (as opposed to the conditional expectations that are hidden in ϱ).

With this understanding of (21), given any measure ν, we define

$$\nu(\phi) = \int_{\mathbb{R}^d} \phi(x)\nu(dx)$$

on the set $C_b(\mathbb{R}^d)$ of all bounded continuous functions ϕ on $\mathbb{R}^d$. We can readily check that Assumptions 3.1–3.3 in Tan and Touzi (2013) are satisfied. Hence, by an application of their main duality result, we can rewrite the KVA as

$$\mathrm{KVA}_0(\varrho) = \inf_{\phi\in C_b(\mathbb{R}^d)} \Big\{\mu_0(\Phi_0) - e^{-(r+h)T}\mu_T(\phi)\Big\}, \tag{22}$$

where the "pseudo-payoff function" ϕ corresponds to a Lagrangian for the constrained optimization problem (21) and where

$$\Phi_0(x) = \sup_{\mathbb{Q}\in\mathcal{Q}(\delta_x)} \mathbb{E}^{\mathbb{Q}}\Big[e^{-(r+h)T}\phi(S_T) + \int_0^T e^{-(r+h)s} hf \sqrt{\frac{d\langle\varrho\rangle}{ds}} ds\Big]. \tag{23}$$

Hence, the KVA in an optimal transportation (OT) setting can be represented as an infimum of KVAs in modified UVM setting.

5.1 Equations in the Markovian Setting

In the Markovian setting of Sect. 3, we consider the probability measures $\mathbb{Q}$ on the canonical space $(\Omega, \mathcal{F}_T)$, under which the canonical process C is a local martingale on $[t, T]$. Define $\mathcal{Q}_t$ as the collection of all such martingale probability measures $\mathbb{Q}$ such that $a_s^{1/2} \in [\underline{\sigma}, \overline{\sigma}]$ $d\mathbb{Q}\times ds$-$a.e.$ on $\Omega\times[t,T]$. Denote $\mathcal{Q}_{t,x} := \{\mathbb{Q}\in\mathcal{Q}_t : \mathbb{Q}[S_s = x, 0\le s\le t] = 1\}$. For any $\phi\in C_b(\mathbb{R}^d)$, let

$$\Phi(t,x) = \sup_{\mathbb{Q}\in\mathcal{Q}_{t,x}} \mathbb{E}^{\mathbb{Q}}\Big[e^{-(r+h)(T-t)}\phi(S_T) + \int_t^T e^{-(r+h)(s-t)} hf \sqrt{\frac{d\langle\varrho\rangle}{ds}} ds\Big], \tag{24}$$

where $\sqrt{\frac{d\langle \varrho \rangle}{dt}} = a_t^{1/2} S_t \left| \partial_S u(t, S_t) - \eta(t, S_t) \right|$, in which u is the solution to (18) with $\lambda = 0$.

Then, in the case where $\eta = (1-\alpha)\partial_S u$, Φ is a viscosity solution to the dynamic programming equation

$$\begin{cases} \Phi(T, S) = \phi(S) \text{ on } (0, \infty) \\ \partial_t \Phi + \sup\limits_{\sigma \in [\underline{\sigma}, \overline{\sigma}]} \left[\mathcal{A}_S^{bs} \Phi + \alpha h f \sigma S |\partial_S u| \right] - (r+h)\Phi = 0 \text{ on } [0, T) \times (0, \infty). \end{cases} \tag{25}$$

In view of (22), in the present OT setup, KVA_0 is obtained as the minimum of

$$\Phi(0, S_0) - e^{-(r+h)T} \int_{\mathbb{R}} \phi(x) \mu_T(dx) \tag{26}$$

over $\phi \in C_b(\mathbb{R}^d)$. This minimization is achieved numerically by the Nelder-Mead simplex algorithm.

As a sanity check, observe that, if μ_T is Black-Scholes σ and $\underline{\sigma} = \overline{\sigma} = \sigma$, then (26) is exactly the time 0 KVA of Sect. 3, independent of ϕ.

6 Numerical Results

Figure 1 shows the results obtained by solving the related PDEs (and minimizing (26) in the OT setup) without model uncertainty as of Sect. 3 (left panel), with UVM uncertainty as of Sect. 4.1 (middle panel) and with OT uncertainty as of

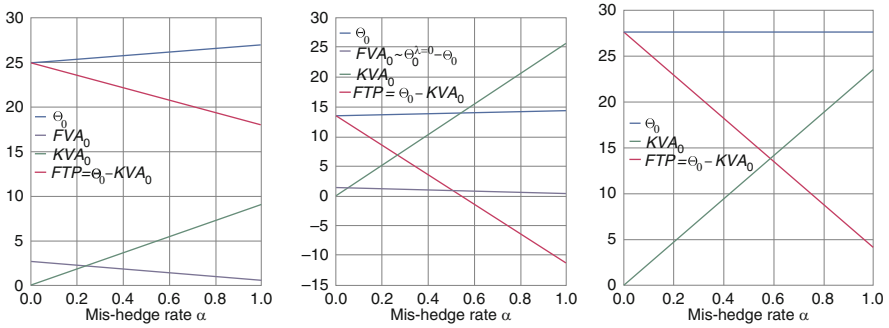


Fig. 1 XVAs and FTP as a function of the mis-hedge parameter α. *Left*: Without model uncertainty. *Middle*: With UVM uncertainty ($\underline{\sigma} = 15\%, \overline{\sigma} = 60\%$). *Right*: With OT uncertainty ($\underline{\sigma} = 15\%, \overline{\sigma} = 60\%, \sigma = 30\%$)

Sect. 5.1 (right panel), for a level of the mis-hedge parameter α increasing from 0 to 100%. We used the following parameters:

$$S_0 = 100, \quad r = 2\%, \quad q = 0, \quad \sigma = 30\%,$$
$$\lambda = 200 \text{ bps}, \quad f = 1.2, \quad h = 10\%$$

and considered a single call option of maturity $T = 5$ years and strike $K = 107$.

The main observation from the left panel is that, unless the hedge is very good (of the order of 25% of mis-hedge or less), the KVA dominates the FVA, and becomes about ten times greater than the FVA in the absence of hedge ($\alpha = 1$). This is logical given that EC has only an indirect reduction effect on the FVA, whereas it directly sizes the KVA.

Going to the middle panel, the FVA changes little, but both u and the KVA (unless the hedge is almost perfect) are tremendously impacted by the uncertainty on the volatility. Regarding the KVA this is in line with the fact that it is the cost of a risk measure, which nonlinearly amplifies the impact of perturbations to its input data.

In reality the time 0 price of a vanilla option such as the one considered in our numerics is given by the market, so there is no model risk on it, but only on the KVA. This is what is reflected by the OT right panel. The model risk on the KVA component however is essentially the same as in the UVM case, because it is conservatively assessed by using the UVM u in (25), fault of a developed theory of valuation at future time points under uncertain volatility subject to the unconditional law of S_T.

XVA desks, KVA in particular, are the first consulted desks in all major trades today. Our results in a toy model where all the quantities of interest can be computed exactly (modulo the numerical error on the PDE solutions) emphasize that, accounting for model risk, the relative importance of the KVA should become even larger. Moreover one can easily imagine how to transpose these results to the setup of Albanese et al. (2016) where each option payoff $(S_{T_i} - K_i)^+$ is replaced by the CVA exposure of the bank to the default at time of its counterparty i, at the (random) time T_i, with corresponding position of the bank $\omega_i S_{T_i}$ and margins received by the bank $\omega_i K_i$. However in this case a relevant risk measure really needs to be computed at a 1-year horizon (as opposed to instantaneous in (6)), in order to leave time to credit events to develop. This points out to developments of a slightly different nature, which would be interesting to develop.

Acknowledgements The research of Stéphane Crépey benefited from the support of the "Chair Markets in Transition", Fédération Bancaire Française, of the ANR project 11-LABX-0019 and of the EIF grant "Collateral management in centrally cleared trading". The travel expenses of Stéphane Crépey regarding its participation to the ICASQF 2016 conference were funded by l'Institut Français de Colombie, Carrera 11 No. 93–12, Ambassade de France en Colombie. The research of Chao Zhou is supported by NUS Grants R-146-000-179-133 and R-146-000-219-112.

References

Albanese, C., Caenazzo, S., Crépey, S.: Capital valuation adjustment and funding valuation adjustment. arXiv:1603.03012 and ssrn.2745909 (2016). Short version "Capital and funding" published in *Risk Magazine* May 2016, 71–76

Avellaneda, M., Levy, A., Paras, A.: Pricing and hedging derivative securities in markets with uncertain volatilities. Appl. Math. Finance **2**(2), 73–88 (1995)

Biagini, S., Bouchard, B., Kardaras, C., Nutz, M.: Robust fundamental theorem for continuous processes. Math. Finance (2015). doi:10.1111/mafi.12110

Kruse, T., Popier, A.: BSDEs with monotone generator driven by Brownian and Poisson noises in a general filtration. Stoch. Int. J. Probab. Stoch. Process. **88**(4), 491–539 (2016)

Peng, S., Yang, Z.: Anticipated backward stochastic differential equations. Ann. Probab. **37**(3), 877–902 (2009)

Soner, H., Touzi, N., Zhang, J.: Wellposedness of second order BSDEs. Probab. Theory Relat. Fields **153**(2), 149–190 (2012)

Tan, X., Touzi, N.: Optimal transportation under controlled stochastic dynamics. Ann. Probab. **41**(5), 3201–3240 (2013)

Index

J.A. Londoño et al. (eds.), *Actuarial Sciences and Quantitative Finance*, Springer Proceedings in Mathematics & Statistics 214, https://doi.org/10.1007/978-3-319-66536-8