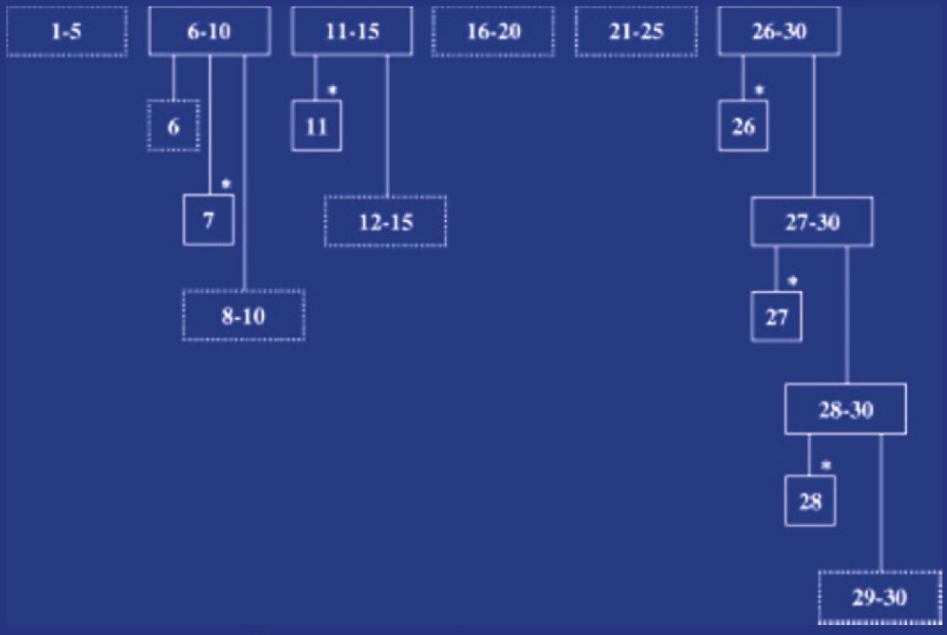
# Screening

## Methods for Experimentation in Industry, Drug Discovery, and Genetics

**Angela Dean**
**Susan Lewis**
Editors




Springer

# Screening

## Methods for Experimentation in Industry, Drug Discovery, and Genetics

Angela Dean
Susan Lewis
Editors

# Screening

## Methods for Experimentation in Industry, Drug Discovery, and Genetics

With 52 Figures

Springer

Angela Dean
Statistics Department
Ohio State University
1958 Neil Avenue
Columbus 43210-1247
U.S.A.

Susan Lewis
School of Mathematics
University of Southampton
Highfield
Southampton SO17 1BJ
United Kingdom

Library of Congress Control Number: 2005929861

ISBN 10: 0-387-28013-8 eISBN: 0-387-28014-6
ISBN 13: 978-0387-28013-4

Printed on acid-free paper.



Printed in the United States of America. (TB/SBA)

9 8 7 6 5 4 3 2 1

springer.com

# Contents


List of Authors vii

Preface xiii

1. An Overview of Industrial Screening Experiments 1
*Douglas C. Montgomery and Cheryl L. Jennings*

2. Screening Experiments for Dispersion Effects 21
*Dizza Bursztyn and David M. Steinberg*

3. Pooling Experiments for Blood Screening and Drug Discovery 48
*Jacqueline M. Hughes-Oliver*

4. Pharmaceutical Drug Discovery: Designing the Blockbuster Drug 69
*David Jesse Cummins*

5. Design and Analysis of Screening Experiments with Microarrays 115
*Paola Sebastiani, Joanna Jeneralczuk, and Marco F. Ramoni*

6. Screening for Differential Gene Expressions from Microarray Data 139
*Jason C. Hsu, Jane Y. Chang, and Tao Wang*

7. Projection Properties of Factorial Designs for Factor Screening 156
*Ching-Shui Cheng*

8. Factor Screening via Supersaturated Designs 169
*Steven G. Gilmour*

9. An Overview of Group Factor Screening 191
*Max D. Morris*

10. Screening Designs for Model Selection 207
*William Li*

11. Prior Distributions for Bayesian Analysis of Screening Experiments 235
*Hugh Chipman*

12. Analysis of Orthogonal Saturated Designs 268
*Daniel T. Voss and Weizhen Wang*

13. Screening for the Important Factors in Large Discrete-Event Simulation Models: Sequential Bifurcation and Its Applications 287
*Jack P. C. Kleijnen, Bert Bettonvil, and Fredrik Persson*

14. Screening the Input Variables to a Computer Model Via Analysis of Variance and Visualization 308
*Matthias Schonlau and William J. Welch*

Index 328

# List of Authors

**Bert Bettonvil** is Assistant Professor in the Department of Information Systems and Management at Tilburg University, The Netherlands. His research interests are in applied statistics, mainly within simulation and information management.
Address: Department of Information Systems & Management/Center for Economic Research, Tilburg University, Postbox 90153, 5000 LE, Tilburg, The Netherlands.
Email: B.W.M.Bettonvil@uvt.nl
Web: http://www.uvt.nl/webwijs/english/show.html?anr=572802

**Dizza Bursztyn** has a PhD in Statistics and works at Ashkelon College. Her research interests are in experimental design, robust design, and computer experiments.
Address: Ashkelon College, Ashkelon, Israel.
Email: dizzal@bezeqint.net

**Jane Chang** is an Assistant Professor in the Department of Applied Statistics and Operations Research at Bowling Green State University. Her research interests are in optimal experimental design, the design and analysis of microarray experiments, and multiple testing in two-level factorial designs.
Address: Department of Applied Statistics and Operations Research, Bowling Green State University, Bowling Green, Ohio 43403, USA.
Email: changj@cba.bgsu.edu
Web: http://www.cba.bgsu.edu/asor/bios/chang.html

**Ching-Shui Cheng** is a Professor of Statistics at the University of California, Berkeley, and a Distinguished Research Fellow at Academia Sinica. His interests are in design of experiments and related combinatorial problems.
Address: Department of Statistics, University of California, Berkeley, CA 94720-3860, USA.
Email: cheng@stat.berkeley.edu

**Hugh Chipman** is Associate Professor and Canada Research Chair in Mathematical Modelling in the Department of Mathematics and Statistics at Acadia University. His interests include the design and analysis of experiments, model selection, Bayesian methods, and data mining.
Address: Department of Mathematics and Statistics, Acadia University, Wolfville, NS, Canada, B4P 2R6.
Email: hugh.chipman@acadiau.ca
Web: http://ace.acadiau.ca/math/chipmanh

**David Cummins** is Principal Research Scientist at Eli Lilly and Company. His interests are in nonparametric regression, exploratory data analysis, simulation, predictive inference, machine learning, model selection, cheminformatics, genomics, proteomics, and metabonomics.
Address: Molecular Informatics Group, Eli Lilly and Company, Lilly Research Laboratories, MS 0520, Indianapolis, IN 46285, USA.
Email: Cummins_DJ@lilly.com

**Angela Dean** is Professor in the Statistics Department at The Ohio State University. Her research interests are in group screening, saturated and supersaturated designs for factorial experiments, and designs for conjoint analysis experiments, in marketing.
Address: Department of Statistics, The Ohio State University, Columbus, OH 43210, USA.
Email: dean.9@osu.edu
Web: http://www.stat.ohio-state.edu/∼amd

**Steven Gilmour** is Professor of Statistics in the School of Mathematical Sciences at Queen Mary, University of London. His interests are in the design and analysis of experiments with complex treatment structures, including supersaturated designs, fractional factorial designs, response surface methodology, nonlinear models, and random treatment effects.
Address: School of Mathematical Sciences, Queen Mary, University of London, Mile End Road, London E1 4NS, UK.
Email: s.g.gilmour@qmul.ac.uk
Web: http://www.maths.qmul.ac.uk/∼sgg

**Jason Hsu** is Professor in the Statistics Department at The Ohio State University. His research interests are in multiple comparisons. Currently he is especially interested in developing methods and software for applications to pharmaceutics and bioinformatics.
Address: Department of Statistics, The Ohio State University, Columbus, OH 43210, USA.
Email: hsu.1@osu.edu
Web: http://www.stat.ohio-state.edu/∼jch

**Jacqueline M. Hughes-Oliver** is Associate Professor of Statistics in the Department of Statistics at North Carolina State University. Her interests are in the analysis of high-dimensional data, pooling experiments, spatial modeling, and design, with applications in drug discovery, ontology-driven analysis of microarray studies, point sources, and transportation modeling.
Address: Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA.
Email: hughesol@stat.ncsu.edu
Web: http://www4.stat.ncsu.edu/∼hughesol

**Joanna Jeneralczuk** is a PhD student in Statistics at the Department of Mathematics and Statistics of the University of Massachusetts in Amherst. Her research concerns the design of experiments, microarray analysis, scoring rules, general information theory, and applications of linear functional analysis to statistics.
Address: Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, USA.
Email: jeneral@math.umass.edu

**Cheryl L. Jennings**, PhD, is Vice-President of Process Design and Improvement, and Six Sigma Black Belt at Bank of America. Her interests are in the application of statistical methods to process design, control, and improvement.
Address: Bank of America, 2727 S 48th St., Tempe, AZ 85282, USA.
Email: cheryl.jennings@bankofamerica.com

**Jack P.C. Kleijnen** is a Professor of Simulation and Information Systems at Tilburg University and a Professor of Operations Research at Wageningen University and Research Centre. His research interests are in the statistical design and analysis of simulation experiments, information systems, and supply chains.
Address: Department of Information Systems & Management/Center for Economic Research, Tilburg University, Postbox 90153, 5000 LE, Tilburg, The Netherlands.
Email: kleijnen@UvT.nl
Web: http://center.kub.nl/staff/kleijnen

**Susan Lewis** is Professor of Statistics in the School of Mathematics at Southampton University and Director of the Southampton Statistical Sciences Research Institute. Her research interests are in group screening, design algorithms, and the design and analysis of experiments in the manufacturing industry.
Address: School of Mathematics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
Email: sml@maths.soton.ac.uk
Web: http://www.maths.soton.ac.uk/∼sml

**William Li** is Associate Professor in the Department of Operations and Management Science at the University of Minnesota. His interests are in experimental design, optimal designs, and quality improvement.
Address: Carlson School of Management, University of Minnesota, Minneapolis, MN 55455, USA.
Email: wli@csom.umn.edu
Web: http://www.csom.umn.edu/∼wli

**Douglas C. Montgomery** is Professor of Engineering and Professor of Statistics at Arizona State University. His research interests are in response surface methodology, empirical modeling, applications of statistics in engineering, and the physical sciences.
Address: Department of Industrial Engineering, Arizona State University, Tempe, Arizona, AZ 85287, USA.
Email: doug.montgomery@asu.edu
Web: http://www.eas.asu.edu/∼masmlab/montgomery

**Max Morris** is a Professor in the Statistics Department and in the Department of Industrial and Manufacturing Systems Engineering at Iowa State University. His research interests include the development and application of experimental designs and strategies for computer simulations, problems involving spatial and dynamic systems, and factor screening experiments.
Address: Department of Statistics, and Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50011, USA.
Email: mmorris@iastate.edu
Web: http://www.public.iastate.edu/∼mmorris

**Fredrik Persson** is Assistant Professor of Production Economics in the Department of Production Economics, Linköping Institute of Technology in Linköping, Sweden. His research interests include modelling and simulation of manufacturing systems and supply chains. Of special interest are simulation methodology and validation methods.
Address: Department of Production Economics, Linköping Institute of Technology, S-581 83 Linköping, Sweden.
Email: fredrik.persson@ipe.liu.se
Web: http://infoweb.unit.liu.se/ipe/fp/presentation_english

**Marco F. Ramoni** is Assistant Professor of Pediatrics and Medicine at Harvard Medical School, Assistant Professor of Oral Medicine, Infection, and Immunity at Harvard School of Dental Medicine, Assistant Professor of Health Sciences Technology at Harvard University and the Massachusetts Institute of Technology, and Associate Director of Bioinformatics at the Harvard Partners Center for Genetics and Genomics. His interests are in Bayesian statistics and artificial intelligence, and their applications to functional and population genomics.
Address: Harvard Partners Center for Genetics and Genomics, Harvard Medical School, New Research Building, Boston, MA 02115, USA.
Email: marco_ramoni@harvard.edu
Web: http://www.hpcgg.org/Faculty/ramoni.jsp

**Matthias Schonlau** is Head of the RAND Statistical Consulting Service. His research interests include computer experiments, data mining, web surveys, and data visualization.
Address: RAND Corporation, 201 N Craig Street, Suite 202, Pittsburgh, PA 15213, USA.
Email: matt@rand.org
Web: http://www.schonlau.net

**Paola Sebastiani** is Associate Professor in the Department of Biostatistics and adjunct Associate Professor in the Bioinformatics and Systems Biology Program of Boston University. Her research interests are in Bayesian data analysis and experimental design, and the automation of statistical methods.
Address: Department of Biostatistics, Boston University School of Public Health, 715 Albany Street, Boston, MA 02118, USA.
Email: sebas@bu.edu.
Web: http://people.bu.edu/sebas

**David M. Steinberg** is an Associate Professor of Statistics in the Department of Statistics and Operations Research at Tel Aviv University. His research interests are in the design of experiments, especially robust design and computer experiments, and in applications in medicine and seismology.
Address: Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel.
Email: dms@post.tau.ac.il
Web: http://www.math.tau.ac.il/∼dms

**Daniel Voss** is Professor of Statistics and Chair of the Department of Mathematics and Statistics of Wright State University. His interests are in design and analysis of experiments and multiple comparisons, with special interest in the analysis of unreplicated factorial experiments.
Address: Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, USA.
Email: dvoss@math.wright.edu
Web: http://www.wright.edu/dan.voss

**Tao Wang** is Assistant Professor in the Department of Epidemiology and Biostatistics at University of South Florida. His research interests are in the design and analysis of microarray experiments and statistical computing.
Address: Department of Epidemiology and Biostatistics, 13201 Bruce B. Downs Blvd., Tampa, FL 33612, USA.
Email: wang.598@osu.edu

**Weizhen Wang** is Associate Professor of Statistics in the Department of Mathematics and Statistics of Wright State University. His interests are in bioequivalence, multiple comparisons, categorical data analysis, quality control, and nonparametric hypothesis testing.
Address: Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, USA.
Email: wwang@math.wright.edu

**William J. Welch** is Professor in the Department of Statistics, University of British Columbia. His research interests include design and analysis of computer experiments, quality improvement, data mining, and statistical methods for drug discovery.
Address: Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, V6T 1Z2, Canada.
Email: will@stat.ubc.ca
Web: http://www.stat.ubc.ca/people/will

# Preface

*Screening* is the process of using designed experiments and statistical analyses to sift through a very large number of features, such as factors, genes or compounds, in order to discover the few features that influence a measured response. In this book, international experts provide accounts of various aspects of screening. They explain and illustrate recent advances and commonly applied methods that are important tools in fields as diverse as industrial quality improvement, engineering research and development, genetic and medical screening, drug discovery, simulation and computer experiments. They also highlight available software for implementing the methods and open issues for future research. The aim of the book is to help practitioners and researchers not only learn about methodologies developed for their own fields, but also to have access to methods from other fields that might usefully be adapted to their own work.

The scene is set for industrial screening in chapter 1 where Montgomery and Jennings describe methods for detecting "active" factors by the use of fractional factorial experiments. They illustrate these methods via a plasma etching investigation from the field of semiconductor manufacturing. In their context, an "active factor" is one which produces different *mean* responses as the level of the factor is changed. The aim is to manipulate the response to a particular target by setting the values, or levels, of the active factors appropriately. This theme is modified by Bursztyn and Steinberg in chapter 2. Here, an "active factor" is one whose levels affect the *variability* of the measured response. Identification of active factors then allows the response variability to be controlled through choice of factor level settings. In modern industrial experimentation, generally the goal is to combine the dual aims of reducing variability and achieving a target mean response, as popularized by Taguchi.

The use of "pooling experiments" began nearly one hundred years ago, initially in dilution studies for estimating the density of organisms in a medium. Pooling experiments today, as described by Hughes-Oliver in chapter 3, are used to make cost savings and gains in precision when investigating large numbers of features. Hughes-Oliver explains how such studies are making a substantial impact in drug discovery as well as in blood screening. The challenges of screening a huge number of chemical compounds during drug development are recounted in more detail

in chapter 4 by Cummins, who describes a wide variety of methods, examples, problems, and experiences from the pharmaceutical industry.

Related to drug discovery is the field of genetic screening for "active genes" linked to the occurrence of a disease. The technology of microarrays and various methods of analysis of data from genetic screening experiments are described in chapter 5 by Sebastiani, Jeneralczuk, Ramoni, and in chapter 6 by Hsu, Chang, and Wang. The former chapter describes methods, both Bayesian and non-Bayesian, for analysing experiments using single-channel synthetic oligoneucleotide microarrys and then presents methods for sample size determination. The latter chapter focuses on 2-channel microarrays and the concerns that arise in simultaneous or multiple testing for active genes.

In the presence of many factors, industrial experiments often need to be designed with fewer observations than can be accommodated in a fractional factorial design. These smaller designs and the analyses of the data sets from the experiments are the concerns of the next three authors. First, Cheng, in chapter 7, discusses the use of designs with complex aliasing for screening. He shows how such designs can have superior *projection properties* so that they provide good information on the small number of active factors and their interactions. Second, Gilmour, in chapter 7 describes supersaturated designs which have even fewer observations than the number of factorial effects to be investigated. He discusses methods of constructing such designs and gives a variety of analysis methods for the resulting experimental data. Third, Morris, in chapter 9, describes a related, but different, approach to reducing the numbers of observations in a screening experiment by investigating factors in groups. This "grouped screening" technique is related to the pooled screening techniques of chapter 3 but in a different practical context.

Li, in chapter 10, turns our attention to designing an experiment when the aim is to select the best response model from a very large set of possible models. Issues of model estimation and discrimination between models are discussed and recommendations for some efficient designs are made. chapters 11 and 12 give more details about methods of analysis of screening experiments. In chapter 11, Chipman describes Bayesian methods for identifying the active factors through the detection of active factorial effects and he illustrates the approach via an experiment in clinical laboratory testing. Voss and Wang, in chapter 12, return to the problem of multiple testing discussed in chapter 6 but in the context of fractional factorial experiments with no degrees of freedom for error. They explain a number of techniques for testing for active effects and describe confidence interval construction for effect sizes.

In some experimental situations, data are obtained by computer generation or simulation rather than through physical experimentation. In such experiments, large numbers of factors can be handled, although it is frequently impossible to obtain large quantities of data due to the time needed for running the computer code. In chapter 13, Kleijnen, Bettonvil, and Persson describe the recent technique of "sequential bifurcation" for finding active factors. They illustrate the method by evaluating three supply chain configurations of varying complexity, studied for an Ericsson factory in Sweden. The techniques described can also be used

for active effect identification from physical experiments. Experiments involving complex computer codes are the concern of Welch and Schonlau in chapter 14. Such experiments may have hundreds of input variables, in which case identification of the important variables is crucial. Methods are described for decomposing a complex input-output relationship into effects which can be easily interpreted and visualized. The methodology is demonstrated on a computer model of the relationship between environmental policy and the world economy.

All of the chapters have been reviewed and we are indebted to the following referees for their help and excellent suggestions: Jane Chang, Hugh Chipman, Jon Forster, Steven Gilmour, Jason Hsu, Jacqueline Hughes-Oliver, Jack Kleijnen, William Li, Wei Liu, Yufeng Liu, Douglas Montgomery, William Notz, Shiling Ruan, Sujit Sahu, Paola Sebastiani, David Steinberg, Anna Vine, Hong Wan, and David Woods.

The idea for this book arose from a Research Section Ordinary Meeting of the Royal Statistical Society and we are grateful to Denise Lievesley, then president of the Society, for suggesting the book. We would like to thank John Kimmel for his guidance and encouragement throughout the preparation process and the Springer production team for their help.

We could not have finished the book without the typing and reference-checking skills of research students from the University of Southampton, UK, especially Roger Gill, Philip Langman, Andrew Rose, and Robert Stapleton. We thank David Woods for his expert help in dealing with figures. We are indebted to our families who supported us throughout this endeavour.

Our most grateful thanks go to all of the authors for their expertise, their time and their patience. We hope that readers of this book will experience as much pleasure as we have in learning about the various techniques.

Angela Dean
Susan Lewis
July 2005

# 1
# An Overview of Industrial Screening Experiments

Douglas C. Montgomery and Cheryl L. Jennings

An overview of industrial screening experiments is presented, focusing on their applications in process and product design and development. Concepts and terminology that are used in later chapters are introduced and explained. Topics covered include a discussion of the general framework of industrial experimentation, the role in those activities played by screening experiments, and the use of two-level factorial and fractional factorial designs for screening. Aliasing in fractional factorial designs, regular and nonregular designs, design resolution, design projection, and the role of confirmation and follow-up experiments are discussed. A case study is presented on factor screening in a plasma etching process from semiconductor manufacturing, including a discussion of the regular fractional factorial design used and the analysis of the data from the experiment.

## 1 Introduction

Statistical experimental design methods are widely used in industry as an important part of the product realization process. Their range of applications includes the design and development of new products, the improvement of existing product designs, evaluation of material properties, and the design, development, and improvement of the manufacturing process. There is also a growing number of applications of designed experiments in business processes, such as finance, marketing or market research, logistics, and supply chain management as discussed in Chapter 13. The objective of most industrial experiments is to characterize the performance of a system (such as a product or a process) and ultimately to optimize its performance in terms of one or more output responses.

The usual framework of industrial experimentation is response surface methodology. First introduced by Box and Wilson (1951), this methodology is an approach to the deployment of designed experiments that supports the industrial experimenter's typical objective cf systems optimization. Myers and Montgomery (2002) described response surface methodology in terms of three distinct steps:

1. Factor screening;
2. Finding the region of the optimum;
3. Determining optimum conditions.

The objective of a factor screening experiment is to investigate, efficiently and effectively, the factors of a system that possibly may be important to its performance and to identify those factors that have important effects. Once the important factors have been identified from the screening experiment, the experimenter will typically move the region of experimentation from the initial location towards one more likely to contain the optimum. The method of steepest ascent (Myers and Montgomery, 2002, Chapter 5) is the procedure most widely employed for this activity. Finally, once near the optimum, the experimenter will usually conduct one or more experiments in order to obtain a fairly precise description of the response surface and an estimate of the optimum conditions.

Response surface methodology is a sequential procedure, with each step consisting of one or more fairly small experiments. Sequential experimentation is important because it makes efficient use of resources and it allows process or system knowledge gained at previous steps to be incorporated into subsequent experiments with the size of each individual experiment remaining small. It is important to keep experiments small because the probability of successfully completing an industrial experiment is inversely proportional to the number of runs that it requires. A widely followed guideline is to allocate no more than 25% of the total available resources to the initial screening experiment. The factor screening step is critical. In many industrial settings, the number of factors that is initially considered to be important can be large (because there is relatively little process knowledge). Eliminating the unimportant factors at an early stage allows the remaining steps of response surface methodology to be completed more quickly, with fewer resources, and usually with a higher overall likelihood of success.

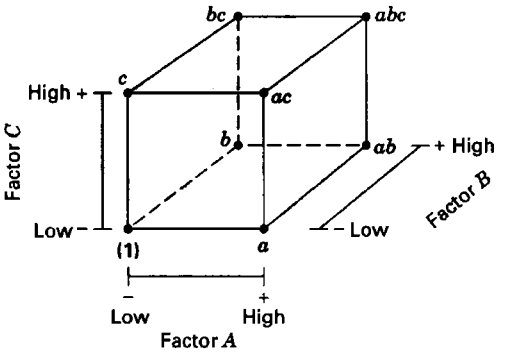
This chapter provides an overview of industrial screening experiments, focusing on process and product design and development. As in any experiment, the pre-experimental planning aspects of a screening experiment are very important. We define pre-experimental planning in terms of the following steps:

1. Problem definition
2. Identifying the response variable(s)
3. Identifying the factors to be studied in the design, including their levels and ranges
4. Selecting the experimental design

Several aspects of these steps are discussed subsequently. For a broader discussion of planning for industrial experiments, see Coleman and Montgomery (1993) and Montgomery (2005), including the supplementary material on the World Wide Web for this text (www.wiley.com/college/montgomery). Other useful references include Andrews (1964), Barton (1997, 1998, 1999), Bishop et al. (1982), and Hahn (1977, 1984).

## 2 Factorial Experiments for Factor Screening

For purposes of factor screening, it is usually sufficient to identify the main effects of the important factors and to obtain some insight about which factors may be




(a) Geometric view

| Run | Factor A | B | C |
|---|---|---|---|
| 1 | – | – | – |
| 2 | + | – | – |
| 3 | – | + | – |
| 4 | + | + | – |
| 5 | – | – | + |
| 6 | + | – | + |
| 7 | – | + | + |
| 8 | + | + | + |

(b) The design matrix

FIGURE 1. A $2^3$ factorial design.

involved in two-factor interactions. Consequently, factorial designs are the basis of most industrial screening experiments. Nearly all of these experiments involve two levels for each of the $f$ factors, so the $2^f$ series of factorial designs are logical factor screening designs.

A $2^3$ factorial design is shown in Figure 1. The geometric view in Figure 1(a) illustrates that each of the $n = 8$ design points can be depicted as residing at the corner of a cube. The two levels of the factors $A$, $B$, and $C$ in the actual units are referred to as low and high. This terminology is used regardless of whether the factors are quantitative or qualitative.

The low and high levels of the factors in the design are expressed as $-1$ and $+1$, respectively, in coded or design units. Sometimes the "1" is omitted, as shown in the design matrix in Figure 1(b) which is written in standard order. In Figure 1(a), the points are labeled with lower-case letters corresponding to factors at their high level. For example, $ac$ refers to $A$ and $C$ at the high level and $B$ at the low level; this is common alternative notation.

The *main effect* of a factor in the $2^f$ system is defined as the average change in response that is observed when the factor is changed from its low level to its high level. Thus the main effect of factor $A$ is the difference in the average response on the right side of the cube in Figure 1(a) and the average response on the left side. The contrasts for calculating all of the effect estimates for the $2^3$ factorial design are shown geometrically in Figure 2. Notice that the main effects of $A$, $B$, and C are the differences in average response on opposite faces of the cube and that the effects of the two-factor interactions $AB$, $AC$, and $BC$ are the differences in averages on planes connecting opposing diagonal edges of the cube. The geometry of the three-factor interaction effect $ABC$ is more complicated and is defined as the difference in average response on two tetrahedrons formed by the eight corners of the cube; see Montgomery (2005) or Box et al. (1978) for a worked example.

The $2^f$ design supports a model containing the $f$ main effects, the $\binom{f}{2}$ two-factor interactions, the $\binom{f}{3}$ three-factor interactions, and so forth, up to and including the

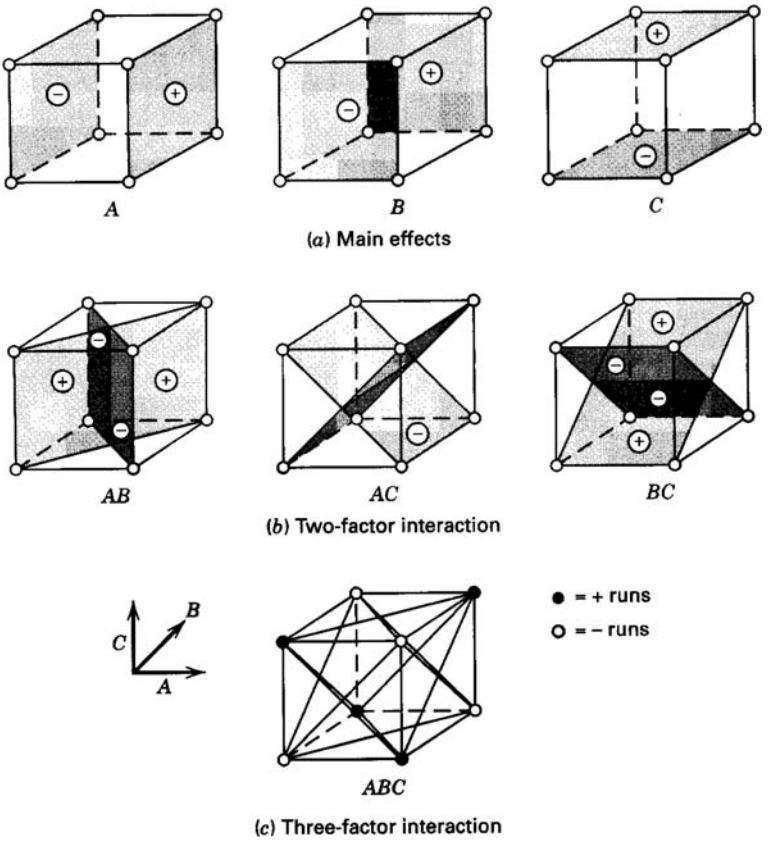



FIGURE 2. Contrasts for calculating the effect estimates in a $2^3$ factorial design.

$f$-factor interaction. For the $2^3$ design of Figure 1, this model is

$$Y = \beta_0 + \sum_{j=1}^{3} \beta_j x_j + \sum_{i<j} \sum_{j=2}^{3} \beta_{ij} x_i x_j + \beta_{123} x_1 x_2 x_3 + \varepsilon, \tag{1}$$

where $Y$ is the response variable, the $x_j$s represent the levels of the design factors expressed in coded units $(-1, +1)$, $\varepsilon$ is the random error component, and the $\beta$s are regression coefficients. It is often a reasonable assumption that the error terms are normally and independently distributed with mean zero and constant variance $\sigma^2$. The least squares estimates of the regression coefficients are exactly one-half of the factor effect estimates defined earlier and the intercept is estimated as the average of all $n$ observations (runs).

When the $2^f$ design is replicated, the statistical significance of the factor effects can be evaluated formally using the analysis of variance. However, in screening experiments with four or more factors, it is relatively common to conduct only

one trial or observation at each design point. This is called an unreplicated or single replicate factorial design. This practice is in keeping with the objective of not running large experiments. An obvious risk when conducting an unreplicated experiment is that we may end up fitting a model to noise; that is, if the response $Y$ is highly variable, the noise component of the observed response may be large relative to the signal and misleading conclusions may result from the experiment.

One way that experimenters can ensure that the noise will not overwhelm the signal in an unreplicated design is to space the low and high levels of the factors aggressively. In a screening experiment, if the experimenter wants to determine the effect of a factor, then the factor must be changed enough to provide a reasonable chance of observing its true effect. A common mistake in industrial screening experiments is to be too conservative in choosing the factor levels. This often leads to disappointing results, when factors shown later to be important are missed in the initial stages of experimentation. Now the experimenter must be careful about selecting the factor levels, and sound engineering or scientific judgment, as well as some experience and practical insight regarding the process, is necessary and should be applied to ensure that unreasonable or potentially dangerous changes in factors are not made. However, remember that in factor screening the mission of the experimenter is often similar to that of Captain James T. Kirk and the crew of the Starship *Enterprise*: "to *boldly* go where no one has gone before" (emphasis added).

The analysis of an unreplicated $2^f$ design is typically conducted using a normal, or half-normal, probability plot of the effect estimates or, equivalently, the estimates of the model regression coefficients. Montgomery (2005) gives details of these probability plots and several examples. The interpretation of these plots is subjective, and some experimenters use more formal analysis procedures, often to support or provide additional guidance regarding the conclusions drawn from the graphical analysis. The method proposed by Lenth (1989) is easy to implement (see Chapter 12) and is beginning to appear in some computer packages (see, for example, Design-Expert, Version 6). The conditional inference chart proposed by Bisgaard (1998–1999) is also a useful supplement to the normal probability plot. Hamada and Balakrishnan (1998) reviewed and compared methods for analyzing unreplicated designs.

A potential concern in the use of a two-level factorial design is the implicit assumption of linearity in the true response function. Perfect linearity is not necessary, as the purpose of a screening experiment is to identify effects and interactions that are potentially important, not to produce an accurate prediction equation or empirical model for the response. Even if the linear approximation is only very approximate, usually sufficient information will be generated to identify important effects. In fact, the two-factor interaction terms in equation (1) do model some curvature in the response function, as the interaction terms twist the plane generated by the main effects. However, because the factor levels in screening experiments are usually aggressively spaced, there can be situations where the curvature in the response surface will not be adequately modeled by the two-factor interaction

terms. In such cases a logical model to consider is the complete second-order model,

$$Y = \beta_0 + \sum_{j=1}^{f} \beta_j x_j + \sum_{i<j} \sum_{j=2}^{f} \beta_{ij} x_i x_j + \sum_{j=1}^{f} \beta_{jj} x_j^2 + \varepsilon. \tag{2}$$

In this model, the regression coefficients of the pure quadratic terms (the $\beta_{jj}$) are not estimable because the typical screening design has all factors at only two levels. However, the experimenter should be alert to the possibility that the second-order model is required. By adding center points to the basic $2^f$ factorial design we can obtain a formal test for second-order curvature, that is, a test of the null hypothesis

$$H_0 : \sum_{j=1}^{f} \beta_{jj} = 0 \text{ versus } H_1 : \sum_{j=1}^{f} \beta_{jj} \neq 0. \tag{3}$$

The center points consist of $nc$ replicates run at the design point $x_j = 0$, $j = 1, 2, \ldots, f$, when all of the design factors are quantitative. Let $\bar{y}_F$ be the average of the response values at the $n_F$ factorial design points and $\bar{y}_C$ be the average response at the center points. The $t$-statistic for testing the null hypothesis in (3) is

$$t_0 = \frac{\bar{y}_F - \bar{y}_C}{\sqrt{MS_E \left( \frac{1}{n_F} + \frac{1}{n_C} \right)}},$$

where $MS_E$ is the mean square for error from the analysis of variance. If the only replication in the design is at the center, then $t_0$ is based on $n_C - 1$ degrees of freedom. Some computer software packages report the $t$-statistic for curvature, whereas others report the $F$-statistic that is the square of $t_0$, and some report both.

When curvature is significant, it will be necessary to include the pure quadratic terms in the model. This requires the experimenter to augment the $2^f$ design (plus center points) with additional runs. The usual choices for augmentation are the $2f$ axial runs $(\pm\alpha, 0, 0, \ldots, 0)$, $(0, \pm\alpha, 0, \ldots, 0)$, $\ldots$, $(0, 0, 0, \ldots, \pm\alpha)$ plus (usually) additional center runs to form a *central composite design*. The axial runs allow the pure quadratic terms in the second-order model to be estimated. Typical choices for $\alpha$ include unity (resulting in a face-centered cube), $\alpha = \sqrt[4]{n_F}$ (resulting in a rotatable design), or $\alpha = f$ (resulting in a spherical design). These choices impart different properties to the central composite design, and are discussed in detail by Khuri and Cornell (1996) and by Myers and Montgomery (2002).

When the screening experiment is conducted in an ongoing process, it is usually a good idea to choose the current operating conditions as the center point. This can provide assurance to the operating personnel that at least some of the runs in the experiment will be conducted under familiar conditions for which a satisfactory product should be manufactured. Runs at these center points could also be used to check for unusual conditions during the execution of the experiment by comparing the response at the center points to historical process performance,

perhaps through use of a control chart. It is also a good idea to space the center points approximately evenly through the randomized run order of the other design points so that an assessment of trend in the response during the experiment may be made. In some screening experiments, there may be little prior information about variability. By running two or three center points as the first runs in the screening experiment, a preliminary estimate of variability can be made. If this estimate seems unreasonably large, then the experiment can be halted until the reasons for the unexpectedly large variability can be determined. Finally, our discussion of center points has focused on the case where all $f$ factors are quantitative. In some screening experiments, there will be at least one qualitative or categorical variable and several quantitative ones. Center points can still be used in these situations by placing them in the centers of the regions involving only quantitative factors.

## 3 Screening Experiments with $2^{f-q}$ Fractional Factorial Designs

Screening experiments often involve a large number of variables (factors). Consequently, many such experiments will require a fractional factorial design. The $2^{f-q}$ fractional factorial design is a logical choice for most factor screening situations. This design is a $1/2^q$ fraction of a $2^f$ design. For example, a $2^{3-1}$ design is a one-half fraction of a $2^3$ design and has four runs, a $2^{6-2}$ design is a one-quarter fraction of a $2^6$ design and has 16 runs, and a $2^{8-4}$ design is a one-sixteenth fraction of a $2^8$ design also with 16 runs. Eight- and sixteen-run fractional factorial designs are used extensively for factor screening.

Table 1 shows a one-half fraction of a $2^4$ design. This design was constructed by first writing down the levels of the "basic design"; that is, a full two-level factorial design that contains the desired number of runs for the fraction. The basic design in Table 1 is a $2^3$ factorial design having 8 runs. Then the levels of the fourth factor are determined by equating one of the interaction columns of the basic design (here $ABC$) to the fourth factor, $D$. The $ABC$ interaction column is found by multiplying columns $A$, $B$, and $C$. The relationship $D = ABC$ is called the design generator. Multiplying both sides of $D = ABC$ by $D$ results in $D^2 = ABCD$. Now $D^2$ is a column of +1s, called the identity column $I$, and $I = ABCD$ is called the defining

TABLE 1. A $2^{4-1}$ fractional factorial design.

| Run | $A$ | $B$ | $C$ | $D = ABC$ |
|---|---|---|---|---|
| 1 | −1 | −1 | −1 | −1 |
| 2 | +1 | −1 | −1 | +1 |
| 3 | −1 | +1 | −1 | +1 |
| 4 | +1 | +1 | −1 | −1 |
| 5 | −1 | −1 | +1 | +1 |
| 6 | +1 | −1 | +1 | −1 |
| 7 | −1 | +1 | +1 | −1 |
| 8 | +1 | +1 | +1 | +1 |

relation for the fractional design. In a one-half fraction the defining relation will contain exactly one word. In the $2^{4-1}$ design of Table 1, this word is $ABCD$. All $2^{f-q}$ fractional factorial regular designs may be constructed by using this general procedure. (Nonregular designs are discussed in Section 5 and also in Chapter 7)

## *3.1 Aliasing in a Fractional Factorial Design*

Because a fractional factorial design uses less than the complete set of factorial runs, not all of the parameters in the complete model supported by the full factorial can be uniquely estimated. Effect estimates are linked together through aliases. For example, in the $2^{4-1}$ design in Table 1, the aliases are

$$\begin{aligned} A &= BCD \quad AB = CD \\ B &= ACD \quad AC = BD \\ C &= ABD \quad BC = AD \\ D &= ABC. \end{aligned}$$

Thus, in this $2^{4-1}$ fractional factorial design, each factorial effect has a single alias; the four main effects are aliased with the four three-factor interactions, and the two-factor interactions are aliased with each other in pairs. Only one effect in each alias chain can be estimated. Aliases can be found by multiplying the effect of interest through the defining relation for the design. For example, the alias of $A$ is found by multiplying $A$ by $I = ABCD$, which produces $A = A^2 = BCD = BCD$, because $A^2$ is the identity column.

The design in Table 1 is called the principal fraction of the $2^{4-1}$ design, and the sign in the generator $D = ABC$ is positive (that is, $D = +ABC$). Another one-half fraction could have been constructed by using $D = -ABC$. This design would have all of the levels in column $D$ of Table 5 reversed. The two one-half fractions can be concatenated to form the complete $2^4$ factorial design.

As another example, the design in Table 2 is a one-sixteenth fraction of the $2^7$ design; that is, a $2^{7-4}$ fractional factorial design. This design was constructed by starting with the $2^3$ as the basic design and adding four new columns using the generators $D = AB$, $E = AC$, $F = BC$, and $G = ABC$. The defining relation is made up of $I = ABD$, $I = ACE$, $I = BCF$, and $I = ABCG$, along with all

TABLE 2. A $2^{7-4}$ fractional factorial design.

| Run | $A$ | $B$ | $C$ | $D = ABC$ | $E = AC$ | $F = BC$ | $G = ABC$ |
|---|---|---|---|---|---|---|---|
| 1 | −1 | −1 | −1 | +1 | +1 | +1 | −1 |
| 2 | +1 | −1 | −1 | −1 | −1 | +1 | +1 |
| 3 | −1 | +1 | −1 | −1 | +1 | −1 | +1 |
| 4 | +1 | +1 | −1 | +1 | −1 | −1 | −1 |
| 5 | −1 | −1 | +1 | +1 | −1 | −1 | +1 |
| 6 | +1 | −1 | +1 | −1 | +1 | −1 | −1 |
| 7 | −1 | +1 | +1 | −1 | −1 | +1 | −1 |
| 8 | +1 | +1 | +1 | +1 | +1 | +1 | +1 |

other words that are equal to the identity column. These are the products of the words $ABD$, $ACE$, $BCF$, and $ABCG$ taken two at a time, three at a time, and four at a time. Thus the complete defining relation is

$$I = ABD = ACE = BCF = ABCG = BCDE = ACDF = CDG = ABEF$$
$$= BEG = AFG = DEF = ADEG = CEFG = BDFG = ABCDEFG.$$

The $2^{7-4}$ design is a *saturated design*; by this we mean that the number of factors, $f = 7$, is one less than the number of runs, $n = 8$. The alias relationships here are somewhat more complicated. Specifically, if we ignore all interactions of order three or higher, each main effect is aliased with a chain of three two-factor interactions:

$$A = BD = CE = FG$$
$$B = AD = CF = EG$$
$$C = AE = BF = DG$$
$$D = AB = CG = EF$$
$$E = AC = BG = DF$$
$$F = BC = AG = DE$$
$$G = CD = BE = AF.$$

Therefore, only the main effects can be estimated using this design, and unique interpretation of these estimates would require the assumption that all two-factor and higher interactions are negligible.

## 3.2 Design Resolution

The *resolution* of a fractional factorial design is a convenient way to describe the alias relationships:

- A resolution III design has at least some main effects aliased with two-factor interactions; so the $2^{7-3}$ design in Table 2 is a resolution III design (often denoted $2^{7-4}_{III}$).
- A resolution IV design has main effects clear of (not aliased with) two-factor interactions, but at least some two-factor interactions are aliased with each other; so the $2^{4-1}$ design in Table 1 is a resolution IV design (often denoted $2^{4-1}_{IV}$).
- A resolution V design has main effects clear of two-factor interactions and two-factor interactions clear of each other; the $2^{5-1}$ design with generator $E = ABCD$ is a resolution V design.

Resolution III and IV designs are used extensively in factor screening because they are usually fairly small designs relative to the number of factors studied and because they can provide the essential information required to do effective screening: identification of important main effects, and some insight regarding potentially important two-factor interactions. Specifically, with $n = 8$ runs, an experimenter can investigate $f = 2$ or 3 factors in a full factorial design, $f = 4$

factors in a resolution IV one-half fraction, and $f = 5$, 6, or 7 factors in a resolution III fractional factorial design. With $n = 16$ runs, one may study $f = 2$, 3, or 4 factors in a full factorial design; $f = 5$ factors in a resolution V one-half fraction; $f = 6$, 7, or 8 factors in a resolution IV fraction; and from 9 to 15 factors in a resolution III fractional factorial design.

Sometimes resolution is insufficient to distinguish between designs. For example, consider the $2^{7-2}$ design. A 32-run resolution IV design can be constructed using $F = ABC$ and $G = BCD$ as the generators. Because this is a resolution IV design, all main effects are estimated clear of the two-factor interactions, but the two-factor interactions are aliased with each other as follows.

$$\begin{aligned} AB &= CF \\ AC &= BF \\ AD &= FG \\ AG &= DF \\ BD &= CG \\ BG &= CD \\ AF &= BC = DG. \end{aligned}$$

Note that this choice of generators results in 15 of the 21 two-factor interactions being aliased with each other across seven alias chains. However, if instead we choose $F = ABCD$ and $G = ABDE$ as the generators, then another resolution IV design results, but in this design the two-factor interactions are aliased with each other as follows.

$$\begin{aligned} CE &= FG \\ CF &= EG \\ CG &= EF. \end{aligned}$$

Thus only 6 of the 21 two-factor interactions are aliased with each other across three alias chains. The second design has the property of *minimum aberration*, which ensures that in a design of resolution $R$ the minimum number of main effects is aliased with interactions involving $R - 1$ factors, the minimum number of two-factor interactions is aliased with interactions involving $R - 2$ factors, and so forth. Tables of the appropriate choice of generators to obtain $2^{f-q}$ fractional factorial designs with maximum resolution and minimum aberration are available in many experimental design textbooks (see Montgomery, 2005, page 305, for example) or can be constructed using widely available software packages such as Minitab and Design-Expert.

The *sparsity of effects principle* (see Box and Meyer, 1986) makes resolution III and IV fractional factorial designs particularly effective for factor screening. This principle states that, when many factors are studied in a factorial experiment, the system tends to be dominated by the main effects of some of the factors and a relatively small number of two-factor interactions. Thus resolution IV designs with main effects clear of two-factor interactions are very effective as screening

designs. However, resolution III designs are also excellent choices, particularly when the number of factors is large relative to the number of factors anticipated to produce important effects.

## 3.3 Design Projection

An important aspect of the success of fractional factorial designs is due to their "projection properties". As an illustration, observe from Table 1 that, if any one of the original four factors $A$, $B$, $C$, $D$ can be eliminated after analysis of the experimental data, then the remaining three factors form a full (unreplicated) $2^3$ factorial design. Furthermore, if two factors are dropped, the remaining two factors form two replicates of a $2^2$ design. Thus elimination of unimportant factors results in a stronger (more informative) experiment in the remaining factors and this can greatly facilitate the practical interpretation of a screening experiment.

All $2^{f-q}$ fractional factorial designs possess projective properties in that they contain within themselves either full factorial (possibly replicated) or smaller fractional factorial designs involving fewer factors than originally investigated. Specifically, a design of resolution $R$ contains full factorial designs in any subset of $R-1$ of the original factors. To illustrate, the $2^{7-4}_{III}$ design in Table 2 will project into a full factorial design in any subset of two of the original seven factors. In general, a $2^{f-q}$ design will project into either a full factorial or a regular fractional factorial design in any subset of $p \leq f-q$ of the original $f$ factors. The subsets of factors providing regular fractional factorial designs as projections are those subsets that do not appear as words in the design's defining relation. Thus, for example, the $2^{7-4}_{III}$ design in Table 2 will project into a full factorial in any subset of three factors that does not form a word in the design's complete defining relation. Therefore the only combinations of factors that, upon projection, will not form a full $2^3$ factorial are $ABD$, $ACE$, $BCF$, $CDG$, $BEG$, $AFG$, and $DEF$.

The projection properties can be extremely useful in the planning stages of a screening experiment if the experimenter has information about the likely importance of the factors. For example, in the $2^{7-4}$ design in Table 2, if the experimenter thought that as many as three of the original factors were likely to be important, it would be a good idea to assign those most likely to be important to a subset of columns that will produce a full factorial upon projection, such as $A$, $B$, and $C$. For more details and recent work on projection properties, see Chapter 7.

## 3.4 Confirmation and Follow-Up Experiments

Interpretation of a fractional factorial experiment always requires careful study of the results, engineering or scientific knowledge about the process being studied, and sometimes the judicious use of Occam's razor.[1] Confirmation experiments

[1] Law of Parsimony used by William of Occam, English philosopher and theologian, c1285–c1349.

should always be conducted to ensure that the experimental results have been interpreted properly. A typical confirmation experiment uses the fitted model from the original experiment to predict the response at a new point of interest in the design space and runs the process at this point to see if the predicted and observed response values are in reasonable agreement. This often provides valuable insight about the reliability of the conclusions that have been drawn.

There are many other strategies for follow-up experimentation after a fractional factorial experiment. These include dropping and adding factors from the original design, rescaling some factors because they were varied over inappropriate ranges in the original experiment, replication of some runs either to improve the precision of estimation of effects or because some runs were not made correctly, shifting the region of experimentation to take advantage of an apparent trend in the observed response and, if center points were used, augmentation to account for apparent curvature.

Not infrequently, we find that the interpretation of a fractional factorial experiment involves some ambiguities. For example, suppose that in the $2^{7-4}$ design in Table 2 we found that the three largest contrast estimates were associated with the main effects of $A$, $B$, and $D$ (and their aliases). The simplest possible interpretation of this finding is that there are three large main effects. However, the alias relationships suggest many other possibilities, such as large main effects $A$, $B$, and their interaction $AB$, or $A$, $D$, and $AD$, or $B$, $D$, and $BD$, and so forth. Unless process knowledge or experience can be used to resolve these ambiguities, additional runs will be necessary. In a resolution III design the simplest way of adding runs is to use the *foldover procedure*. To fold over a resolution III design, simply run another experiment that is the "mirror image" of the first, that is, with all signs reversed. The combined design consisting of the original design plus its foldover form a resolution IV experiment from which all main effects can be estimated clear of the two-factor interactions. This procedure is widely used in practice, particularly with eight-run resolution III screening designs. A somewhat less widely used variation of the full foldover is to fold over on a single column (change the signs within the single column), which will allow the main effect of that single factor and all two-factor interactions involving that factor to be estimated clear of other two-factor interactions.

It is also possible to fold over resolution IV designs, but the procedure is more complicated. Because the main effects are already estimated clear of the two-factor interactions, an experimenter folds over a resolution IV design for reasons such as (1) to break as many two-factor interaction alias chains as possible, (2) to separate the two-factor interactions on a specific alias chain, or (3) to isolate one or more specific two-factor interactions. Montgomery and Runger (1996) discussed and illustrated these issues and provided a table of recommended foldover arrangements for resolution IV designs with $6 \leq f \leq 10$ factors. In our experience, however, a complete foldover of a resolution IV design is often unnecessary. Generally, there are one or two (or a very few) aliased interactions that need to be identified. These interactions can usually be de-aliased by adding a smaller number of runs to the original fraction than would be required by a full foldover. This technique is sometimes called a partial foldover or semifolding. Montgomery (2005, pages 329–331)

presented a complete example of the procedure. Other useful references include Mee and Peralta (2000), Nelson et al. (2000), Li and Mee (2002), and Li and Lin (2003).

## 4 An Example of an Industrial Screening Experiment

Plasma etching is a widely used process in semiconductor manufacturing. A fractional factorial design was used to study the effect of six factors, labeled $A$, $B$, $C$, $D$, $E$, on a measure of how uniformly the wafer has been etched, namely, the range of thickness measurements over the entire wafer. Based on experience with similar etching processes, the experimenters felt that the main effects of the six factors and some of the two-factor interactions involving these factors were likely to be important, but that interactions involving three or more factors could be safely ignored. They were also comfortable with the assumptions of normality and constant variance for the response. This led to selection of the $2^{6-2}_{IV}$ design with $E = ABC$ and $F = BCD$ in Table 3. This table shows the six design factors and the observed responses that resulted when the experiment was conducted. The complete defining relation and the alias relationships for main effects, two-factor and three-factor interactions for this design are shown in Table 4.

Figure 3 shows a half-normal probability plot of the effect estimates obtained from the Design-Expert software package. Three main effects, $A$ (pressure), $B$ (power), and $E$ (gap) are important. Because the main effects are aliased with three-factor interactions, this interpretation is probably correct. There are also two two-factor interaction alias chains that are important, $AB = CE$ and $AC = BE$. Because $AB$ is the interaction of two strong main effects, those of pressure and

TABLE 3. The design and data for the plasma etching screening experiment.

| Run order | Factor A: Pressure torr | Factor B: Power watts | Factor C: He sccm | Factor D: CF4 sccm | Factor E: Gap cm | Factor F: Thickness Angstroms | Response: Range Angstroms |
|---|---|---|---|---|---|---|---|
| 15 | −1 | −1 | −1 | −1 | −1 | −1 | 441 |
| 5 | 1 | −1 | −1 | −1 | 1 | −1 | 289 |
| 1 | −1 | 1 | −1 | −1 | 1 | 1 | 454 |
| 8 | 1 | 1 | −1 | −1 | −1 | 1 | 294 |
| 12 | −1 | −1 | 1 | −1 | 1 | 1 | 533 |
| 13 | 1 | −1 | 1 | −1 | −1 | 1 | 100 |
| 2 | −1 | 1 | 1 | −1 | −1 | −1 | 405 |
| 6 | 1 | 1 | 1 | −1 | 1 | −1 | 430 |
| 3 | −1 | −1 | −1 | 1 | −1 | 1 | 427 |
| 7 | 1 | −1 | −1 | 1 | 1 | 1 | 329 |
| 14 | −1 | 1 | −1 | 1 | 1 | −1 | 469 |
| 11 | 1 | 1 | −1 | 1 | −1 | −1 | 392 |
| 10 | −1 | −1 | 1 | 1 | 1 | −1 | 558 |
| 4 | 1 | −1 | 1 | 1 | −1 | −1 | 112 |
| 16 | −1 | 1 | 1 | 1 | −1 | 1 | 436 |
| 9 | 1 | 1 | 1 | 1 | 1 | 1 | 373 |

TABLE 4. Defining relation and aliases for the plasma etching experiment in Table 3.

$$I = ABCE = BCDF = ADEF$$

| | |
|---|---|
| $A = BCE = DEF$ | $AB = CE$ |
| $B = ACE = CDF$ | $AC = BE$ |
| $C = ABE = BDF$ | $AD = EF$ |
| $D = AEF = BCF$ | $AE = BC = DF$ |
| $E = ABC = ADF$ | $AF = DE$ |
| $F = ADE = BCD$ | $BD = CF$ |
| | $BF = CD$ |

$$ABD = ACF = BEF = CDE$$
$$ABF = ACD = BDE = CEF$$

power, and $BE$ is the interaction between power and gap, it is likely that these are the proper interpretations of these two-factor interactions. An analysis of variance for the reduced model containing only these main effects and interactions is shown in Table 5.

The model equation (in coded units) that results from this experiment is

$$\hat{y} = 377.61 - 87.75x_1 + 28.95x_2 + 51.78x_5 + 53.39x_1x_2 - 26.91x_2x_5,$$

where $x_1$, $x_2$, and $x_5$ are the levels of $A$, $B$, and $E$, respectively, and $\hat{y}$ is the

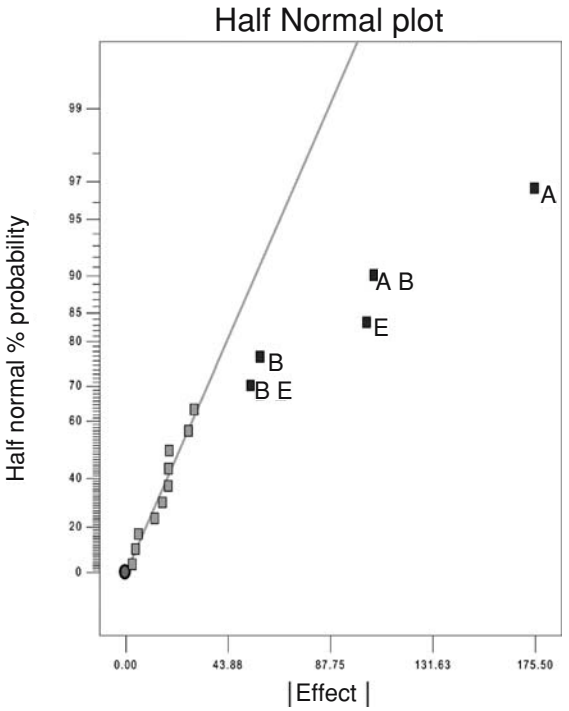



FIGURE 3. Half-normal probability plot of the effect estimates from the plasma etching experiment in Table 3.

TABLE 5. Analysis of variance for the reduced model identified in the plasma etching screening experiment.

| Source of variation | Sum of squares | Degrees of freedom | Mean square | $F$ Value | Prob > $F$ |
|---|---|---|---|---|---|
| Model | 236693.20 | 5 | 47338.64 | 37.48 | <0.0001 |
| *A* | 123201.00 | 1 | 123201.00 | 97.53 | <0.0001 |
| *B* | 13409.64 | 1 | 13409.64 | 10.62 | 0.0086 |
| *E* | 42890.41 | 1 | 42890.41 | 33.95 | 0.0002 |
| *AB* | 45603.60 | 1 | 45603.60 | 36.10 | 0.0001 |
| *BE* | 11588.52 | 1 | 11588.52 | 9.17 | 0.0127 |
| Residual | 12631.62 | 10 | 1263.16 | | |
| Corrected Total | 249324.80 | 15 | | | |

predicted response. A normal probability plot of the studentized residuals (see Montgomery, 2005, page 397) from this model, shown in Figure 4, indicates that there are no problems with the normality assumption and that there are no outliers. Plots of the studentized residuals versus the fitted values and the design factors did not reveal any problems with inequality of variance.

In the plasma etching experiment, the response variable is the range of thickness measurements on the wafer after etching. It is a measure of the consistency or

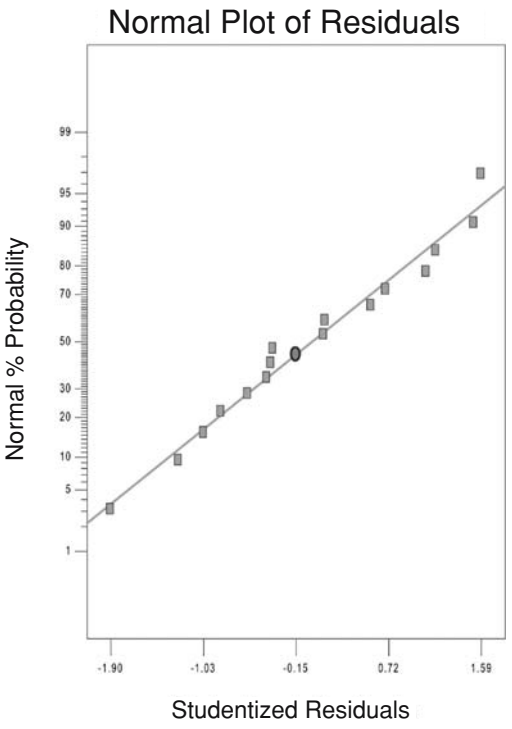



FIGURE 4. Normal probability plot of the studentized residuals from the plasma etching experiment.

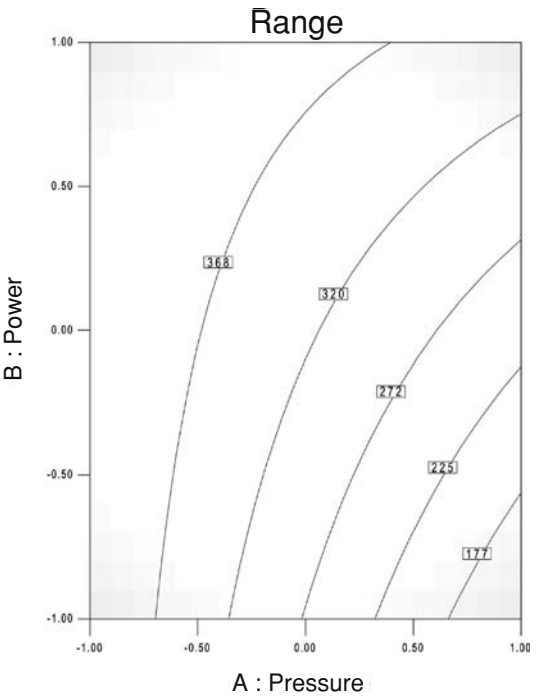



FIGURE 5. Contour plot for the thickness range (Angstroms) in terms of Power (watts) and Pressure (torr) for the plasma etching experiment with gap at the low level.

uniformity of the etching process over the entire wafer. Therefore, small values of the response are desirable. Figure 5 shows a contour plot of the response as a function of pressure and power, with gap set at the low level. (The other design factors have no impact on this contour plot because they are not included in the model.) This plot shows the region where small responses are observed. The indication is that operating conditions with low power, high pressure, and low gap will result in small values of the response variable. Additional experiments should be conducted in the vicinity of these settings to confirm these findings. Other follow-up experiments might include the determination of whether there is a region of lower settings for power and gap and higher pressure that will result in further decreases in the range of thickness, or the fitting of a second-order model to obtain a more precise estimate of the location of the optimum operating conditions.

## 5 Some Other Aspects of Industrial Screening Experiments

In this section, we briefly discuss a few other topics on screening experiments that occasionally arise in practice.

The $2^{f-q}$ fractional factorial designs that we discussed in Section 3 all require that the number of runs in the design is a power of two. There are times where this

may seem inconvenient; for example, if there are $f = 7$ factors, then the saturated fractional factorial design in Table 2 requires only $n = 8$ runs, but the addition of one more factor leads to a fractional factorial design with $n = 16$ runs, or twice as many runs. Plackett and Burman (1946) developed a family of two-level resolution III fractional factorial designs where the number of runs is only required to be a multiple of four.

When $n$ is a power of two, the Plackett–Burman designs are identical to the usual $2^{f-q}$ regular fractional factorial designs. However, the Plackett–Burman designs for $n = 12, 20, 24, 28, 32, \ldots$, are potentially useful for reducing the number of runs in a factor screening experiment. For example, if there are $f = 8, 9, 10$, or 11 factors, the Plackett–Burman design would require only 12 runs whereas the conventional $2^{f-q}$ fractional factorial design has 16 runs. When $n$ is a multiple of four that is not a power of two, the Plackett–Burman design is called a nongeometric design. These designs have very complex alias structures. To illustrate, for the Plackett–Burman design with $f = 11$ factors and $n = 12$ runs, every main effect is partly aliased with every two-factor interaction that does not involve that main effect. Thus each of the 11 main effects is partly aliased with 45 two-factor interactions. Furthermore, the two-factor interactions occur in more than one alias chain. For example, $AB$ occurs in every main effect alias chain except those for $A$ and $B$. This is called *partial aliasing*. If there are large two-factor interaction effects, partial aliasing can make the interpretation of a Plackett–Burman design difficult. This is a potential disadvantage to the practical deployment of these designs. There are some analysis procedures that could be used for designs with complex aliasing. For example, Hamada and Wu (1992) proposed a stepwise regression-based procedure that under some conditions can simplify the interpretation of a Plackett–Burman design. (For further information on Plackett–Burman designs, see Chapter 7.)

The use of supersaturated designs is another approach for further reducing the number of runs in a factor screening experiment. These are designs where the number of runs is less than the number of factors. Supersaturated designs were proposed by Booth and Cox (1962) and then largely ignored until about a dozen years ago, when researchers began to develop computationally intensive algorithms for their construction. Lin (2000) provided a review of recent developments (see also Chapter 8). The primary potential application of supersaturated designs is in systems with either very large numbers of factors or in situations where each run is very resource intensive as, for example, in many types of computer experiments or scale model tests, such as finite element analysis models of complex structures, wind tunnels, or towing basins.

The analysis of a supersaturated design is usually conducted by using some type of sequential model-fitting procedure, such as stepwise regression. Abraham et al. (1999) and Holcomb et al. (2003) have studied the performance of analysis methods for supersaturated designs. Techniques such as stepwise model fitting and all-possible-regressions type methods may not always produce consistent and reliable results. Holcomb et al. (2003) showed that the performance of an analysis technique in terms of its type I and type II error rate can depend on several factors,

including the type of supersaturated design used, the number of factors studied, and the number of active factors. Generally, a supersaturated design would have the best chance of working satisfactorily in a screening situation when the sparsity of effects principle holds and when the number of runs is at least equal to half the number of factors studied; see also Chapter 8.

Occasionally a situation occurs where the experimenter considers conducting a screening experiment where some of the design factors have more than two levels. A common situation is that some factors have two levels and others have three levels. There are many fractional factorial designs with "mixed" levels that could be used in these situations. However, these designs generally have moderately complex alias relationships, and this can lead to some difficulties in practical interpretation. If factors are quantitative, the only reason to use more than two levels in a screening experiment is to account for (or guard against) potential curvature. Generally, keeping all factors at two levels and adding center points to the design is a safer experimental strategy, although in some situations it could lead to a slightly larger experimental design. Trading a small increase in the size of the experiment to achieve simplicity in design execution and data interpretation is usually an important consideration.

Special care needs to be taken in screening experiments when one or more of the factors is qualitative or categorical. For example, catalyst type (organic versus nonorganic) is a categorical factor. When categorical factors are present, we often find that main effects and interactions between some quantitative factors may be different at each level of the categorical factors. This can result in some unanticipated high-order interactions. For example, if the time–emperature interaction is strong but not consistent for the two types of catalyst, then a three-factor interaction is present. If the experimenter has selected a design assuming that all three-factor interactions are negligible (a common assumption in screening experiments), then misleading conclusions may result. Multi-level categorical factors are especially problematic as this increases the likelihood that high-order interactions may be present. It is sometimes helpful to conduct separate experiments at each level of the categorical factor to reduce the complexity of the problem. When there are several categorical factors, screening can be very difficult because it is not uncommon to find that the system only responds satisfactorily to certain combinations of these categorical factors, and that other combinations produce no useful information about the response. This is essentially a system that is dominated by (possibly high-order) interactions rather than main effects. Breaking the overall screening exercise into a series of smaller experiments is a strategy that can be effective in some of these situations.

## *References*

Abraham, B., Chipman, H., and Vijayan, K. (1999). Some risks in the construction and analysis of supersaturated designs. *Technometrics*, **41**, 135–141.

Andrews, H. P. (1964). The role of statistics in setting food specifications. *Proceedings of the Sixteenth Annual Conference of the Research Council of the American Meat Institute*,

43–56. Reprinted in *Experiments in Industry: Design, Analysis, and Interpretation of Results*. Editors: R. D. Snee, L. B. Hare and J. R. Trout. American Society for Quality Control, Milwaukee, 1985.

Barton, R. R. (1997). Pre-experiment planning for designed experiments: Graphical methods. *Journal of Quality Technology*, **29**, 307–316.

Barton, R. R. (1998). Design-plots for factorial and fractional factorial designs. *Journal of Quality Technology*, **30**, 40–54.

Barton, R. R. (1999). *Graphical Methods for the Design of Experiments*, Springer Lecture Notes in Statistics 143. Springer-Verlag, New York.

Bisgaard, S. (1998–1999). Conditional inference chart for small unreplicated two-level factorial experiments. *Quality Engineering*, **11**, 267–271.

Bishop, T., Petersen, B., and Trayser, D. (1982). Another look at the statistician's role in experimental planning and design. *The American Statistician*, **36**, 387–389.

Booth, K. H. and Cox, D. R. (1962). Some systematic supersaturated designs. *Technometrics*, **4**, 489–495.

Box, G. E. P. and Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics*, **28**, 11–18.

Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society B*, **13**, 1–45.

Box, G. E. P., Hunter, W. G., and Hunter, S. J. (1978). *Statistics for Experimenters*, John Wiley and Sons, New York.

Coleman, D. E. and Montgomery, D. C. (1993). A systematic approach to planning for a designed industrial experiment (with discussion). *Technometrics*, **35**, 1–27.

Design-Expert Software. Version 6 User's Guide (2000). Stat-Ease Inc., Minneapolis.

Hahn, G. J. (1977). Some things engineers should know about experimental design. *Journal of Quality Technology*, **9**, 13–20.

Hahn, G. J. (1984). Experimental design in a complex world. *Technometrics*, **26**, 19–31.

Hamada, M. and Balakrishnan, N. (1998). Analyzing unreplicated factorial experiments: A review with some new proposals (with discussion). *Statistica Sinica*, **8**, 1–41.

Hamada, M. and Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, **24**, 130–137.

Holcomb, D. R., Montgomery, D. C., and Carlyle, W. M. (2003). Analysis of supersaturated designs. *Journal of Quality Technology*, **35**, 13–27.

Khuri, A. I. and Cornell, J. A. (1996). *Response Surfaces: Designs and Analyses*, second edition. Marcel Dekker, New York.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, **31**, 469–473.

Li, H. and Mee, R. W. (2002). Better foldover fractions for resolution III $2^{k-p}$ designs. *Technometrics*, **44**, 278–283.

Li, W. and Lin, D. K. J. (2003). Optimal foldover plans for two-level fractional factorial designs. *Technometrics*, **45**, 142–149.

Lin, D. K. J. (2000). Recent developments in supersaturated designs, Chapter 18. In *Statistical Process Monitoring and Optimization*. Editors: S. H. Park and G. G. Vining, pages 305–319. Marcel Dekker, New York.

Mee, R. W. and Peralta, M. (2000). Semifolding $2^{k-p}$ designs. *Technometrics*, **42**, 122–134.

Montgomery, D. C. (2005). *Design and Analysis of Experiments*, sixth edition. John Wiley and Sons, New York.

Montgomery, D. C. and Runger, G. C. (1996). Foldovers of $2^{k-p}$ resolution IV designs. *Journal of Quality Technology*, **28**, 446–450.

Myers, R. H. and Montgomery, D. C. (2002). *Response Surface Methodology: Process and Product Optimization using Designed Experiments*, second edition. John Wiley and Sons, New York.

Nelson, B. J., Montgomery, D. C., Elias, R. J., and Maass, E. (2000). A comparison of several design augmentation strategies. *Quality and Reliability Engineering International*, **16**, 435–449.

Plackett, R. L. and Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, **33**, 305–325.

# 2
# Screening Experiments for Dispersion Effects

Dizza Bursztyn and David M. Steinberg

Reduction of the variability in performance of products and manufacturing processes is crucial to the achievement of high levels of quality. Designed experiments can play an important role in this effort by identifying factors with *dispersion effects*, that is, factors that affect performance variability. Methods are presented for the design and analysis of experiments whose goal is the rapid screening of a list of candidate factors to find those with large dispersion effects. Several types of experiments are considered, including "robust design experiments" with noise factors, and both replicated and unreplicated fractional factorial experiments. We conclude that the effective use of noise factors is the most successful way to screen for dispersion effects. Problems are identified that arise in the various analyses proposed for unreplicated factorial experiments. Although these methods can be successful in screening for dispersion effects, they should be used with caution.

## 1 Introduction

The use of statistical techniques to reduce the variability of manufactured products has been a key feature of the quality movement during the last 20 years. In particular, the quality engineering ideas of Genichi Taguchi (1986) stimulated widespread use of designed experiments to reduce variability. The goal of these experiments is to find settings of *design* or *control factors*, whose nominal settings can be controlled in the process or product specification, that improve the quality of the final product. In particular, these experiments have been used in practice to identify design factors that affect process variability. Such factors are said to have *dispersion effects*. Knowledge about dispersion effects can be used to select nominal factor values that reduce variation, thereby designing quality into products. The goal of this chapter is to describe methods that have been proposed for screening a set of candidate factors to find those which have important dispersion effects.

We distinguish between three basic paradigms that have been proposed for identifying and estimating dispersion effects:

1. Include replicate observations at each of the design factor combinations to permit estimation of variance;

2. Force variation into the experiment by including so-called *noise factors* in the experiment;
3. Analyze dispersion from experiments with no replication.

The inclusion of noise factors is one of the unique and important contributions of Taguchi. We illustrate the important role that noise factors can play and explore the relative advantages of the different paradigms in Section 2. The first paradigm, simple replication, leads largely to standard designs and analyses, and this approach is treated only briefly in Section 3 of this chapter. We discuss methods for analyzing experiments with noise factors in Section 4. Much of the research in this area has focused on what can be accomplished with no replication and these methods are presented in Section 5. We discuss some examples in Section 6 and summarize the ideas in Section 7.

Experiments for reducing variability are often called *robust design experiments*, emphasizing the goal of making the product or process insensitive, or robust, to variation in manufacturing or use environments. Interested readers who would like more background on robust design in general can refer to a number of books or review articles. In particular, Phadke (1989) has an excellent discussion of the engineering aspects of robust design. An up-to-date statistical view that exploits a response surface approach to robust design experiments is given in Chapters 10 to 12 of Wu and Hamada (2000) or Chapter 11 of Myers and Montgomery (2002), and the review articles by Steinberg (1996), Montgomery (1999), and Ankenman and Dean (2001).

## 2 Design Strategies

One of the principal questions that arises is how to design an experiment in order to identify dispersion effects. The direct and obvious answer is to include replicate observations at the different design factor combinations used in the experiment. It is then possible to compute sample variances at each point and to use them as the basis for modeling how the dispersion depends on the experimental factors. The most immediate drawback to replication is that it increases the size and cost of the experiment.

An important insight of Taguchi was that performance variation is often caused by natural variation in important input factors. Typical examples include the natural variation of a component part dimension about its nominal value or of a field condition at the time a product is used. Taguchi proposed that these variations should be expressly included as *noise factors* in the experiment. A simple example will clarify the idea and show how such an experiment differs from one with simple replication. Suppose an experiment is run on a product with a component part that has a nominal width of 3 mm. The parts are purchased from a supplier and have an average width equal to the nominal setting and a standard deviation of 0.05 mm. In a design with replication, the actual part used to build

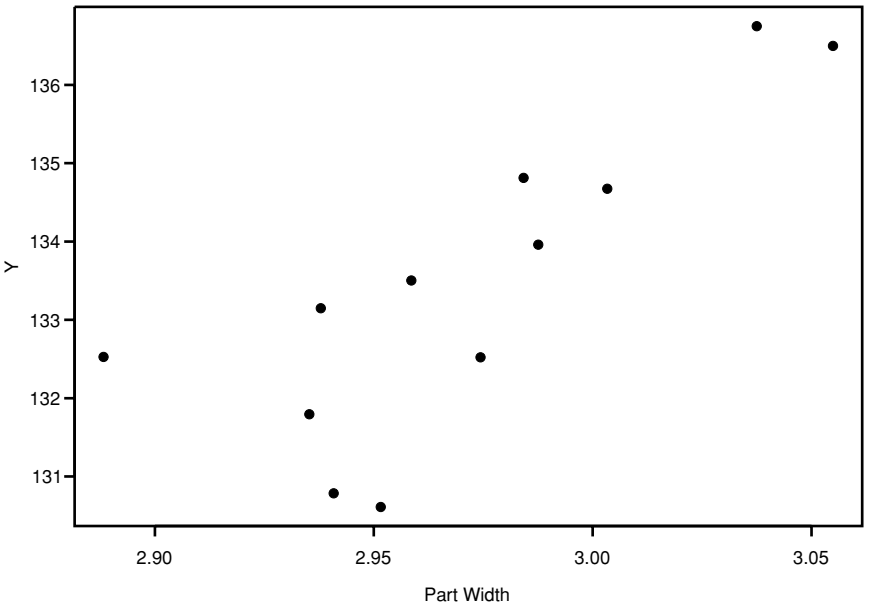



FIGURE 1. The natural variation in Part Width (mm) is seen to be a primary cause for the variation in the performance measure *Y*.

each experimental product would be sampled from the parts in stock, so that some would have widths as low as 2.9 mm, others as high as 3.1 mm, and many would be close to 3.0 mm. Taguchi's idea was to replace the random sample of parts by a directed sample in which only those with widths, say, 2.9 and 3.1 mm, would be used. The deviation from the nominal is always $-0.1$ mm or $+0.1$ mm. The experimental design would include a two-level noise factor for the deviation from the nominal width and a part of width 2.9 or 3.1 mm would be chosen for each experimental run in accordance with the levels of that noise factor as specified by the design.

There can be significant advantages to the inclusion of a noise factor in a robust design experiment. Suppose that the noise factor affects performance. In the example above, this would mean that performance differs as a function of the actual part width. Then the natural variation of the part width transmits variation to the performance. Typical data that might arise from a random sample of manufactured parts are shown in Figure 1. The measured part width (mm) is plotted on the horizontal axis and performance, $y$, on the vertical axis. An experiment in which the width of the part is ignored and only performance is measured (simple replication) will not show the full picture in Figure 1. Instead, it exposes only the marginal spread on the vertical axis. Moreover, simple replication makes no effort to balance the actual widths used at different design factor combinations. Including the noise

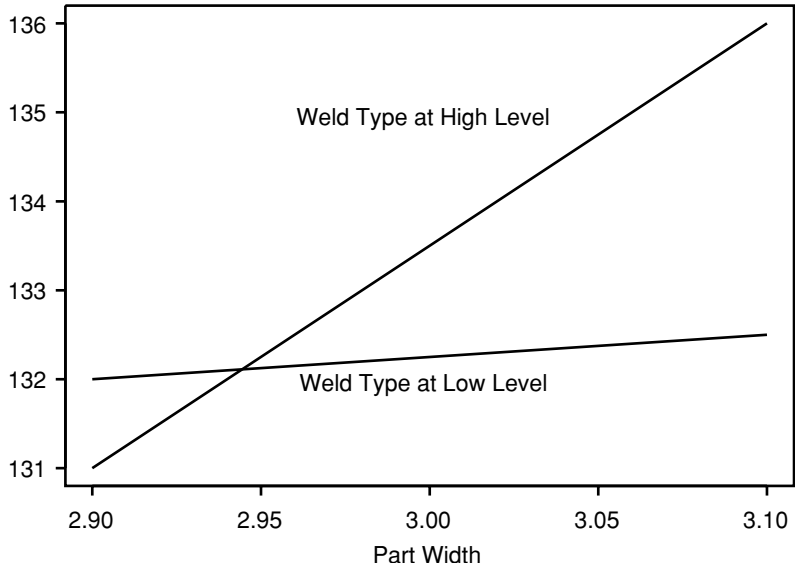



FIGURE 2. Interaction plot of the design factor Weld Type with the noise factor Part Width (mm) with performance measure on the vertical axis. The slope with respect to Part Width is flatter when the Weld Type is at its low level, implying that the low setting of Weld Type will reduce transmitted variation.

factor explicitly in the design removes both disadvantages. It guarantees balance and enables us to make the plot in Figure 1. We can then assess the contribution of the noise factor in terms of its regression slope.

The relation of noise factors to dispersion effects can now be understood by making a simple interaction plot for a noise factor and a design factor, as in Figure 2. The figure shows that the regression slope of performance on the noise factor (part width) differs according to the level of the design factor (weld type). The "flatter" slope occurs when the design factor is at the lower of its two experimental levels, resulting in lower transmitted variation. The design factor has a dispersion effect related to its interaction with the noise factor. The fact that the dispersion effect is found by comparing regression slopes is the key to the enhanced power provided by including noise factors. The relation of interactions to dispersion effects was first emphasized by Shoemaker et al. (1991). Steinberg and Bursztyn (1994, 1998) and Bérubé and Nair (1998) proved that including noise factors in an experiment can significantly enhance its power to detect dispersion effects.

There are also two possible disadvantages to experiments with noise factors. First, it will often be difficult or expensive to control noise factors in an experiment. Secondly, the inclusion of noise factors can substantially increase the size of the experiment, without adding to the number of design factors studied.

# 3 Experiments with Replication

Standard statistical methods can be used for designs that include simple replication. The data at each experimental condition can be summarized by measures of location and spread of the response, typically the sample average $\bar{Y}$ and standard deviation $s$. Models can then be fitted that relate the measure of spread to the experimental factors. If normality is reasonable, one can fit a generalized linear model (McCullagh and Nelder, 1989, Myers et al. 2002) to the standard deviations with the main effects of the factors, and possibly some interactions, as explanatory variables. Vining and Myers (1990) gave a detailed and readable description of how to fit joint response surface models to the averages and standard deviations. See also Box (1988) and Nair and Pregibon (1988).

# 4 Experiments with Noise Factors

## *4.1 Statistical Analysis*

In a robust design experiment with noise factors, each experimental observation is characterized by the levels of the design factors and the noise factors. Two different approaches have been advocated for the analysis of experiments with noise factors: response model analysis and performance measure analysis.

In response model analysis, a single statistical model is fitted to the data, including effects for both the design factors and the noise factors. As explained in Section 2, design factors that interact with noise factors have dispersion effects. So it is important to include such interactions as candidate terms in the analysis. Denoting the design factors by $D_1, \ldots, D_k$ and the noise factors by $N_1, \ldots, N_t$, the model for the $i$th observation with all linear main effects and design by noise interactions is

$$Y_i = \beta_0 + \sum_j \beta_j d_{i,j} + \sum_u \alpha_u n_{i,u} + \sum_{j,u} \alpha_{u,j} d_{i,j} n_{i,u} + \epsilon_i, \tag{1}$$

where $d_{i,j}$ and $n_{i,u}$ give the settings of $D_j$ and $N_u$, respectively, on the $i$th experimental run, $\beta_j$ is the regression coefficient for the main effect of $D_j$, $\alpha_u$ is the regression coefficient for the main effect of $N_u$, and $\alpha_{u,j}$ is the regression coefficient for the interaction of $D_j$ and $N_u$. The effect of the noise factor $N_u$ for given settings $d_1, \ldots, d_k$ of the design factors can be found from (1) by grouping together all the terms that multiply $n_u$. This gives the regression slope

$$\alpha_u(d_1, \ldots, d_k) = \alpha_u + \sum_j \alpha_{u,j} d_j. \tag{2}$$

The experiment provides estimates of the parameters $\alpha_u, \alpha_{u,1}, \ldots, \alpha_{u,k}$ which can be substituted into (2) and the $d_j$ can then be chosen so as to make the noise factor regression slopes as close to 0 as possible. If there is a single design factor that interacts with a noise factor, then interaction plots such as Figure 2 expose the full

details about the dispersion effect. If factors are involved in several interactions, the model equations are the most useful guide to identifying design factor settings that will reduce dispersion. See Shoemaker et al. (1991), Myers et al. (1992), Steinberg and Bursztyn (1994, 1998), Steinberg (1996), and Myers and Montgomery (2002, Chapter 11) for more detail on the response model analysis.

Performance measure analysis was proposed by Taguchi (1986) and his influence on robust design experiments is no doubt a major reason for its continuing popularity. The data at each design factor combination are summarized by a "performance measure", which is analyzed for dependence on the design factors. Settings are then chosen to optimize the performance measure. Taguchi proposed a number of different performance measures, depending on the goals of the experiment and the nature of the data. Of most importance for dispersion effects is his so-called signal-to-noise ratio for experiments whose goal is to reduce variation about a target value, $SN_q = 20 \log(\bar{Y}_q / s_q)$, where $\bar{Y}_q$ and $s_q$ are the average and standard deviation of the response at the $q$th design factor combination. Design factors are identified as having dispersion effects if they have a strong effect on the signal-to-noise ratio. Taguchi did not advocate any formal statistical criteria for distinguishing between strong and weak effects on the signal-to-noise ratio and many published case studies simply pick out the design factors with the largest percent sum of squares in the analysis of variance breakdown of the signal-to-noise ratio. In most of these studies, the design factors exhaust most of the degrees of freedom, so any formal criterion needs to perform effectively in settings with few error degrees of freedom.

Subsequent research has pointed out a number of serious drawbacks to performance measure analysis. Box (1988) showed that the use of the signal-to-noise ratio would be preferable to the standard deviation itself if, overall, the standard deviation were roughly proportional to the average. Such dependence might be anticipated if the data followed lognormal distributions. Box showed that a similar analysis could then be obtained by taking logs of the original data and using the standard deviation as a performance measure. He also showed that the signal-to-noise ratio could be an inefficient measure of dispersion in other settings.

A fundamental criticism of performance measure analysis is that it makes no explicit use of the noise factors in the experiment. The first stage of the analysis is essentially to "collapse" the data over the noise factor settings. The analysis then proceeds in exactly the same manner that would be used for experiments with simple replication. Taguchi's main argument in favor of this approach seems to be simplicity. However, having made efforts to include noise factors in the experiment, one may wonder why they would be ignored in the analysis. Steinberg and Bursztyn (1994) showed that the performance measure analyses may misidentify dispersion effects. They analyzed an experiment in which two design factors both interact with the same noise factor. Although the response model analysis quickly identified the dispersion effects of these factors, the performance measure analysis found, instead, a phantom dispersion effect for the design factor that is aliased with the interaction of the two original design factors. They explained why the erroneous conclusion is built into the performance measure analysis.

Steinberg and Bursztyn (1998) developed additional theory to support the conclusions of their earlier article. They also compared the power to detect a dispersion effect in several types of robust design experiments. There are two important conclusions from their article. First, they found that explicit inclusion of noise factors in an experiment significantly increases the power to detect a dispersion effect relative to experiments with simple replication. Second, they found that the increase in power is obtained only via the response model analysis. The performance measure analysis does not enjoy the same gains in power.

Bérubé and Nair (1998) reached similar conclusions. They also found that the response model analysis was much more efficient than the performance model analysis. In addition, their work highlighted the importance of choosing noise factors that account for a substantial fraction of overall process variation.

## *4.2 Designing an Experiment with Noise Factors*

Most robust design experiments involving noise factors have followed the *cross-product array* format proposed by Taguchi. In this type of experiment, separate experimental arrays are prepared for the design and the noise factors, typically with the smallest sample size that can accommodate the number of factors in each set. Then the two arrays are "crossed" to generate a matrix of design points, with rows corresponding to design factor combinations and columns to noise factor combinations.

The design array might be a two-level fraction, a three-level fraction, or a mixed orthogonal array such as $L18$, which can include up to seven factors at three levels and one more with two levels (see Wu and Hamada, 2000, Chapters 6 and 7, for details on orthogonal array designs). In all the examples that we have seen, noise factors were limited to two levels. Moreover, to reduce the overall sample size, the noise factors are sometimes combined into "compound noise factors," in which an entire set of noise factors may be modified in unison (see also Chapter 9). The compounding technique is most often used when changes in the noise factors have a predictable effect on the direction of the outcome, say in switching from easy to severe use conditions.

The crossed array designs have two important properties, related to the methods of analysis described above. First, these designs enable independent estimation of all "design by noise factor" interactions; that is, interactions involving one design factor and one noise factor (Shoemaker et al., 1991). Thus they are well suited to the response model analysis. Second, they provide a "fair comparison" of the design factor combinations, by subjecting them to identical noise conditions.

Shoemaker et al. (1991) pointed out that crossed array designs are fractional factorial designs in the full set of design and noise factors with a particular aliasing pattern. They suggested treating the design problem from the outset in terms of finding a design with a desirable aliasing pattern. They called the resulting designs *combined arrays* and presented some settings in which a combined array might be preferred to a crossed array. For example, with 4 design factors and 1 noise factor,

all at two levels, the crossed array could be a $2^{4-1}$ fractional factorial design in the design factors run at each of the two levels of the noise factor. Instead, one could run a $2^{5-1}$ design in all 5 factors. Both designs permit estimation of all main effects and design by noise factor interactions. The latter design also permits estimation of design by design factor interactions.

Borkowski and Lucas (1997) further explored the construction of combined array designs. The important contribution in their approach was to set up a framework for the different requirements on design resolution for different types of effects in robust design experiments. They proposed the concept of mixed resolution to characterize separately the resolution of the design for effects involving only design factors, only noise factors, or interactions between design and noise factors.

Another general class of designs that has been proposed for robust design experiments is known as *compound orthogonal arrays*. These designs are similar to crossed arrays in that a fractional factorial design in one set of factors is crossed with levels of the remaining factors. However, rather than repeat the same fraction in the remaining factors, different (equivalent) fractions can be run at each experimental point in the first array. The $2^{5-1}$ design that we described earlier is an example of a compound array, in which the two different $2^{4-1}$ fractions involving the four design factors are run at the two different levels of the noise factor. Compound arrays can be used to improve the aliasing properties of the full design. Rosenbaum (1994, 1996) provided initial properties of compound arrays and Hedayat and Stufken (1999) added considerable detail.

One of the issues that concerned Rosenbaum (1994, 1996) was the derivation of conditions on the design that would guarantee that a performance measure provides an unbiased estimate of the variance across all possible noise conditions. He used this argument as a justification for compound arrays. As we have already written, the performance measure analysis is at best inefficient and at worst misleading, so we find this argument irrelevant in assessing the benefits of compound arrays.

## 5 Unreplicated Factorial Experiments

We now consider factorial experiments in which there is no replication of design factor combinations and no use of noise factors. The idea of identifying dispersion effects in unreplicated factorials again has roots in the work of Taguchi. It was first studied in detail by Box and Meyer (1986) and has since attracted considerable interest and research.

At first glance, one might think that an unreplicated experiment is hopelessly overextended for finding dispersion effects. The standard analysis of an unreplicated design begins by studying the effects of the factors and their interactions on location. The location effects involve orthogonal contrasts that typically exhaust *all* the degrees of freedom in an unreplicated design. The crucial insight of Box and Meyer (1986) was that, often, only a small fraction of the factors really have

substantial location effects and, when projected onto those active factors, the resulting design might then have replication that could be used to study effects on dispersion. (See Chapters 1 and 7 for discussion about projections.)

In subsequent years many papers have been devoted to this topic. A variety of methods has been proposed and some important reservations and caveats have been discovered. More research is being done and more is needed. At the present time, it appears that the initial optimism generated by Box and Meyer (1986) has been substantially moderated. One can, indeed, identify dispersion effects from unreplicated designs, but there are numerous pitfalls and opportunities for misleading conclusions.

Below we describe the most important methods that have been proposed and review some of their pros and cons. Before proceeding, we remark that these same ideas could also be used in a response model analysis of an experiment with noise factors. These methods could be useful if the design factors affect the variance of the random error term $\epsilon_i$ in equation (1).

In many of the descriptions, we refer to "statistics for the dispersion effect of factor $j$". By this we mean either a main effect associated directly with a single factor or an interaction effect associated with some collection of factors. Much of the presentation focuses on two-level factorial designs and $+1$ and $-1$ are used to denote the high and low levels of each factor in these designs.

## *5.1 Dyestuffs Example*

We illustrate the different methods by applying them to a $2^6$ experiment on dyestuffs presented by Box and Draper (1987). We consider here the results for hue, one of three response variables studied in the experiment, and label the six factors $A$ to $F$. In this section we examine the full set of experimental data. However, many screening experiments are smaller in size. So, in the next section, we extract some fractions from the experiment that will be typical of small screening studies.

Figure 3 shows a half-normal plot for hue from the full experiment. Two effects clearly stand out and these are the main effects of factors $A$ and $F$. All the other effects appear consistent with a null hypothesis of no effect. These conclusions are reinforced by other analyses. Fitting a model with all main effects and two-factor interactions results in highly significant effects for factors $A$ and $F$. No other effects are significant at the 5% level, but the main effect of $B$ and the $AC$ interaction are both quite close, with $p$-values less than 0.075. Analysis by Lenth's (1989) method, discussed in Chapter 12, also finds that the only significant effects are those for $A$ and $F$.

A plot of the residuals versus the fitted values from the location model with $A$ and $F$ only is shown in Figure 4. The spread of the residuals is seen to increase with the fitted value, a feature that might be explained by dispersion effects of the experimental factors, in particular those in the location model. We examine this conjecture with the various methods that have been suggested for screening dispersion effects.

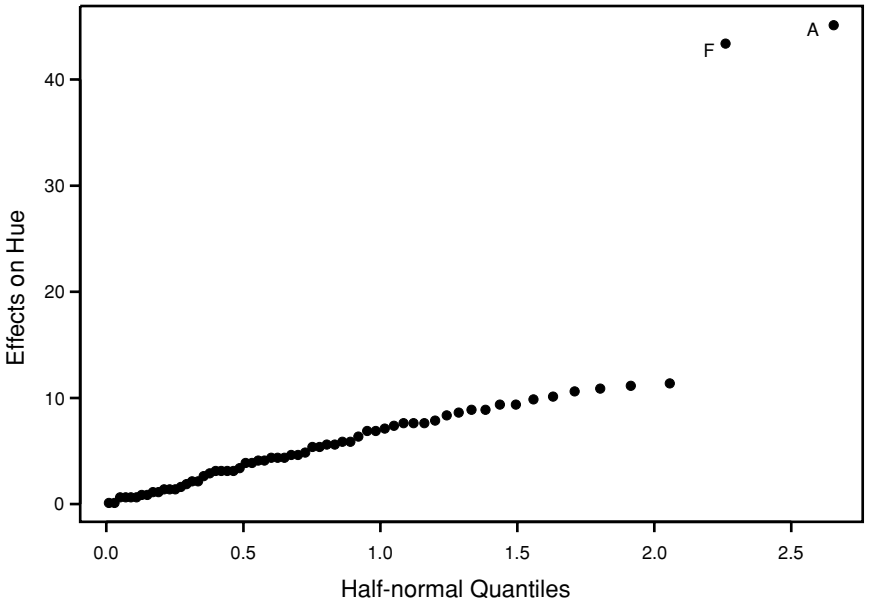



FIGURE 3. Half-normal probability plot of the factor effects from the $2^6$ factorial experiment on hue.

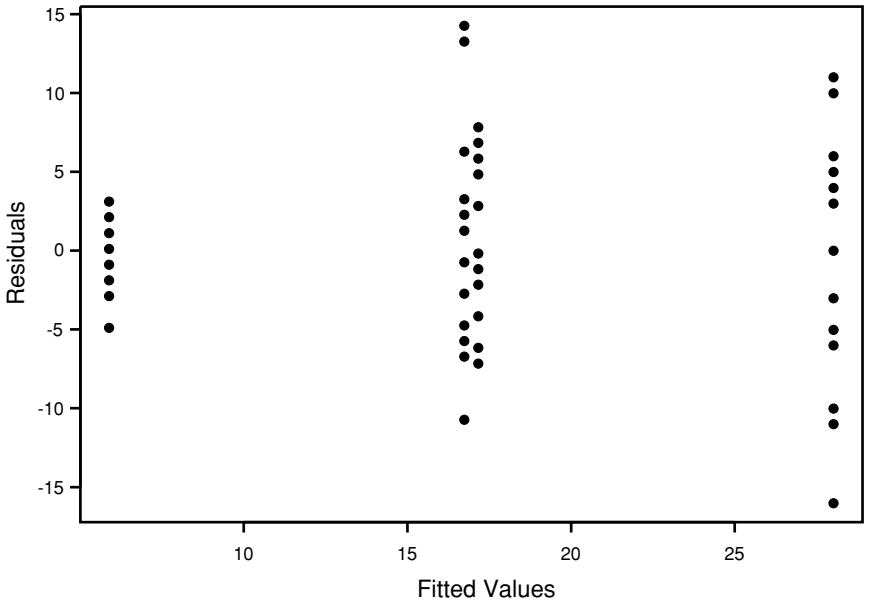



FIGURE 4. Plot of residuals versus fitted values from the $2^6$ experiment on hue. The location model includes the main effects of factors $A$ and $F$ only.

## 5.2 Box and Meyer's Method

Box and Meyer (1986) considered the identification of dispersion effects in unreplicated $2^{k-p}$ designs. The first step in their approach is to identify and estimate the active location effects. Let $r_i (i = 1, \ldots, n)$ denote the residuals from the fitted location model. To examine whether factor $j$ has a dispersion effect (or, equivalently, whether there is a dispersion effect associated with the $j$th main effect contrast), compute the sums of squared residuals at the two levels of this factor:

$$SS(j+) = \sum_{j(+)} r_i^2,$$
$$SS(j-) = \sum_{j(-)} r_i^2.$$

The indices $j(+)$ and $j(-)$ in the above definitions are used to denote summation over the design points with factor $j$ at its high and low levels, respectively. The test statistic proposed by Box and Meyer was one-half the log of the ratio of these sums of squares,

$$D_j^{BM} = 0.5 \log[SS(j+)/SS(j-)]. \tag{3}$$

Box and Meyer did not present any formal inference procedures for using this statistic to identify dispersion effects. The use seems to be informal for screening factors with large effects from those with no or small effects on dispersion, for example, by making a normal probability plot of the statistics; see Montgomery (1990) for an application of this idea.

Box and Meyer also derived a useful result (which is applied in some of the subsequent methods in this chapter) that relates dispersion effects to location effects in regular $2^{k-p}$ designs. We present the result first for $2^k$ designs and then explain how to extend it to fractional factorial designs. First, fit a fully saturated regression model, which includes all main effects and all possible interactions. Let $\hat{\beta}_i$ denote the estimated regression coefficient associated with contrast $i$ in the saturated model. Based on the results, determine a location model for the data; that is, decide which of the $\hat{\beta}_i$ are needed to describe real location effects. We now compute the Box–Meyer statistic associated with contrast $j$ from the coefficients $\hat{\beta}_i$ that are *not* in the location model. Let $i \circ u$ denote the contrast obtained by elementwise multiplication of the columns of $+1$s and $-1$s for contrasts $i$ and $u$. The $n$ regression coefficients from the saturated model can be decomposed into $n/2$ pairs such that for each pair, the associated contrasts satisfy $i \circ u = j$; that is, "contrast $i \circ u$ is identical to contrast $j$". Then Box and Meyer proved that equivalent expressions for the sums of squares $SS(j+)$ and $SS(j-)$ in their dispersion statistic are

$$SS(j+) = (2/n) \sum (\hat{\beta}_i + \hat{\beta}_u)^2, \tag{4}$$

$$SS(j-) = (2/n) \sum (\hat{\beta}_i - \hat{\beta}_u)^2, \tag{5}$$

where the sums extend over all pairs for which $i \circ u = j$ and any regression coefficients that are used in the location model are replaced by 0.

For $2^{k-p}$ fractional factorial designs, the sums of squares can again be written, as in equations (4) and (5), in terms of squared sums and differences of regression

coefficients that are not included in the location model. Depending on the generators used to construct the fractional design, the relevant pairs of contrasts may be those for which $i \circ u = -j$, in which case the expressions for $SS(j-)$ and $SS(j+)$ in (4) and (5) may be switched.

Brenneman and Nair (2001) and McGrath and Lin (2001) showed that the method of Box and Meyer (1986) can be problematic as described below when more than one factor has a dispersion effect. Brenneman and Nair (2001) examined the consequences of using the Box–Meyer method when there is a log-linear dispersion model in which

$$\mathrm{Var}(Y_i) = \sigma_i^2 = \exp(\phi_0 + Z_i^{'}\phi), \tag{6}$$

where $\phi_0$ is a constant, the vector $Z_i$ specifies the levels of the $q$ explanatory variables in the dispersion model, and $\phi$ is a vector of coefficients of the effects. If two factors $A$ and $B$ have dispersion effects in this model, then they also affect the Box–Meyer statistic $D_j^{BM}$ in (3), for the contrast associated with their interaction. This could result in the spurious identification of a dispersion effect for the interaction contrast or in the cancellation of a legitimate dispersion effect.

When applied to the data on hue in the $2^6$ dyestuffs experiment, the Box–Meyer statistic (3) points to factor $F$ as having the most potential for a dispersion effect, with a statistic of 0.59. The effect for factor $F$ stands out, though not dramatically, on a normal plot of the Box–Meyer statistics (Figure 5). The next strongest

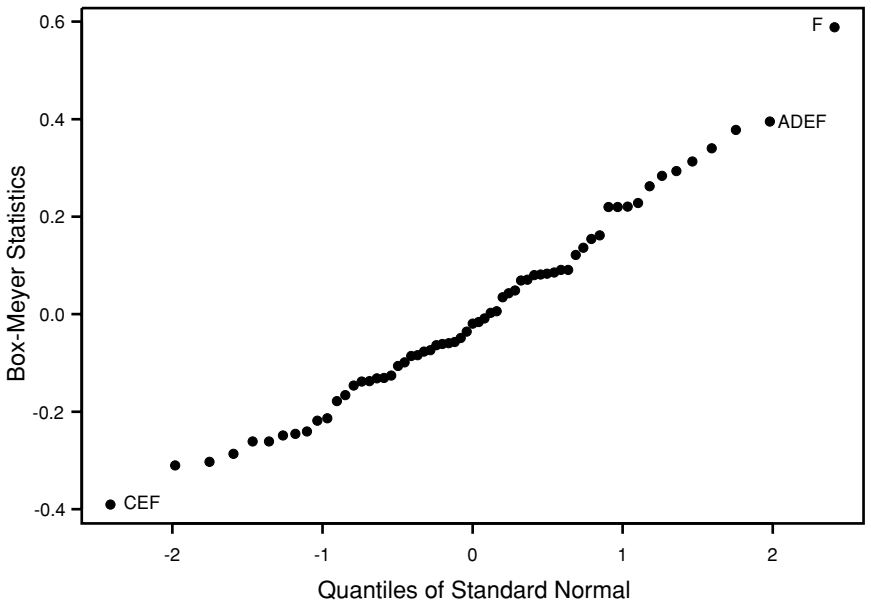



FIGURE 5. Normal probability plot of the Box–Meyer dispersion statistics for the $2^6$ experiment on hue, with $A$ and $F$ in the location model.

effects, both with absolute values of 0.39, are the *ADEF* interaction and the *CEF* interaction. The next strongest main effect is for factor *B*, with a statistic of −0.30.

## 5.3 Bergman and Hynén's Method

Bergman and Hynén (1997) developed a method similar to that of Box and Meyer (1986), but with a simple and exact distribution theory for inference from the test statistic. The important observation of Bergman and Hynén was that the residuals from the fitted location model could complicate inference for the Box–Meyer statistic in two ways. First, the residuals in the two sums of squares could be correlated. Second, the residuals at the high (low) level of factor $j$ typically depend on the actual variances at *both* levels of the factor, not just the level at which the run was made.

Bergman and Hynén proposed a clever solution that corrects both of the problems cited above. The source of both problems is that residuals, being linear combinations of the observations will, in general, include observations at both levels of factor $j$, inducing correlation and making the expected mean square of the residual depend on both variances. Both of these drawbacks can be overcome if the residuals at the high (low) level of factor $j$ depend only on the observations made at that same level. One natural way to accomplish this "separation" of the observations is to fit separate location models at the two levels of $j$. Equivalently, for $2^{k-p}$ designs, one can fit a single, expanded, location model to the full set of data that includes the main effect of $j$ and the product of $j$ with each effect currently in the model. For example, suppose the location model includes the main effects of factors *A* and *B*. Table 1 shows the expanded location models that would need to be fitted to test for dispersion effects associated with factor *A*, factor *C*, or with the *AC* interaction. As is evident in this example, different location models need to be fitted depending on the potential for having a dispersion effect of the factor under consideration. The idea of the expanded location model is also discussed very briefly by Box and Meyer (1986, Section 5).

The Bergman and Hynén (1997) statistic for factor $j$ is given by

$$D_j^{BH} = SS(j+)/SS(j-), \tag{7}$$

TABLE 1. Expanded location models for the Bergman–Hynén method for several potential dispersion effects when the location model includes the main effects of factors *A* and *B*. The location model always includes main effects for *A* and *B* and also includes the main effect of the dispersion candidate and its interactions with *A* and with *B*

| Dispersion term | Terms in location model |
|---|---|
| *A* | *A*, *B*, *AB* |
| *C* | *A*, *B*, *C*, *AC*, *BC* |
| *AC* | *A*, *B*, *AC*, *C*, *ABC* |

where

$$SS(j+) = \sum_{j(+)} r(j)_i^2,$$
$$SS(j-) = \sum_{j(-)} r(j)_i^2,$$

and $r(j)_i$ is the residual for the $i$th observation from the expanded location model for factor $j$. Under the null hypothesis of no dispersion effects, $D_j^{BH}$ has an $F$ distribution with $m(j)$ degrees of freedom in both numerator and denominator, where $m(j) = 0.5(n - b(j))$, $n$ is the total number of observations, and $b(j)$ is the number of location effects in the expanded model for factor $j$. The sums of squares in the Bergman–Hynén statistic, like those in the Box–Meyer statistic, can be expressed in terms of the coefficients from a saturated regression model. The relevant expressions for the Bergman–Hynén statistic are identical to equations (4) and (5), using squared sums and differences of pairs of regression coefficients. The Bergman–Hynén statistic replaces by 0 any term $(\hat{\beta}_i + \hat{\beta}_u)$ or $(\hat{\beta}_i - \hat{\beta}_u)$ that contains coefficients involved in the location model, whereas the Box–Meyer statistic sets only the location model coefficients to 0.

Blomkvist et al. (1997) extended the Bergman–Hynén method to identify dispersion effects from unreplicated multi-level experiments. Arvidsson et al. (2001) showed how the method could be applied to split-plot experiments.

Although the Bergman–Hynén statistic provides a clever correction to some problems with the Box–Meyer statistic, it remains problematic in the face of multiple dispersion effects (see Brenneman and Nair, 2001 and McGrath and Lin, 2001). If factor $j$ alone has a dispersion effect, the numerator and denominator of the statistic $D_j^{BH}$ in (7) are unbiased estimators of the variances at the high and low levels of $j$. However, if several factors have dispersion effects, one has instead unbiased estimates of the *average* variances at these two levels, where the averaging includes the effects of all the other dispersion effects. This dependence of $D_j^{BH}$ on additional dispersion effects can lead to inflated type I error probabilities and thus to spurious identification of dispersion effects.

On the dyestuffs example, the Bergman–Hynén method also signals factor $F$ as being related to dispersion. With the main effects of $A$ and $F$ in the location model, $F$ has a Bergman–Hynén statistic of 3.27 ($p$-value = 0.001). The next strongest effects, as with the Box–Meyer method, are the *ADEF* interaction, with a statistic of 2.24 ($p$-value = 0.017) and the *CEF* interaction, with a statistic of 0.45 ($p$-value = 0.035).

## 5.4 Harvey's Method

Harvey's (1976) method was developed for general regression settings with heteroscedasticity and was aimed more at improving inference for the location parameters than out of intrinsic interest in the dispersion effects. Harvey, like Box and Meyer, begins by fitting a location model. Harvey's idea was to proceed by using ordinary least squares regression to fit a log-linear model to the squared

residuals,

$$\log(r_i^2) = \phi_0 + Z_i'\phi + v_i, \tag{8}$$

where $\phi_0$ and $\phi$ are defined as in equation (6), $Z_i$ is the vector of explanatory variables on the $i$th run for the factors that affect dispersion, and $v_i$ is an error term. For two-level orthogonal designs, the estimated dispersion effect $\hat{\phi}_j$ for factor $j$ is

$$\begin{aligned} D_j^H &= \left(\sum\nolimits_{j(+)} \log(r_i^2) - \sum\nolimits_{j(-)} \log(r_i^2)\right) \Big/ n \\ &= \log\left(\frac{\prod_{j(+)} r_i^2}{\prod_{j(-)} r_i^2}\right)^{1/n}. \end{aligned}$$

The statistic $D_j^H$ is similar to the Box–Meyer statistic $D_j^{BM}$, but uses the geometric averages of the squared residuals rather than the arithmetic averages.

Assuming a log-linear dispersion model as in equation (8), Nair and Pregibon (1988) showed that $D_j^{BM}$ is the maximum likelihood estimator of $\phi_j$ for the model in which factor $j$ is the only one with a dispersion effect, whereas $D_j^H$ is the maximum likelihood estimator of $\phi_j$ for a fully saturated dispersion model with effects for all possible factors. Nair and Pregibon concluded from this result that $D_j^H$ would be a better statistic to use for initial analyses aimed at identifying possible dispersion effects.

Brenneman (2000) found that Harvey's method could underestimate the dispersion effect of factor $j$ if that factor was left out of the location model. This result led Brenneman and Nair (2001) to propose a modified version of Harvey's method for two-level factorial experiments that is based on the results of Bergman and Hynén (1997). In the modified version, the dispersion statistic for factor $j$ is computed from residuals from an expanded location model that includes the effect of factor $j$ and all its interactions with other effects in the location model. For two-level designs, the modified Harvey's statistic for factor $j$ is then

$$D_j^{MH} = \left(\sum_{j(+)} \log(r(j)_i^2) - \sum_{j(-)} \log(r(j)_i^2)\right) \Big/ n.$$

Brenneman and Nair (2001) noted that $D_j^{MH}$ can also sometimes give biased results for an effect that is aliased with the interaction of two factors that have dispersion effects.

An additional problem with Harvey's method is that factorial experiments may yield some residuals equal to zero. Ad hoc modification to Harvey's method is necessary in the presence of any zero residuals. One such suggestion was made by Ferrer and Romero (1995), who replaced any zero residuals by one-half the smallest (nonzero) absolute residual in the experiment. Ferrer and Romero (1993, 1995) also suggested adding 1.27 to the logarithm of each squared residual to correct for the bias in estimating $\phi_0$.

The dyestuffs example illustrates the potential for problems with Harvey's method. With only $A$ and $F$ in the location model, the $ABDEF$ interaction clearly stands out with the largest value of the statistic $D_j^{MH}$. However, subsequent

examination shows that this occurs only because of two zero residuals that are computed (both by S-Plus 2000 and by MATLAB) as nonzero numbers on the order of $10^{-15}$. As these residuals are not zero, they do not trigger an error message in applying Harvey's method, but they completely dominate the dispersion statistics computed. Thus the residuals must be checked as an essential preliminary step for this method. Moreover, standard residual displays do not highlight zero residuals, so it is important to specifically print out any residuals with absolute values below a set threshold. The problem of zero residuals can be even more serious with the modified Harvey's method, as the presence of zero residuals may depend on the particular location model being fitted.

## 5.5 *Wang's Method*

Wang (1989) considered the problem of identifying dispersion effects using data from a two-level orthogonal design of the type discussed in Chapter 1. He assumed that the variance follows a log-linear model (as in our equation (6)) and based his method on testing the null hypothesis that all of the dispersion effects are zero. Let $Z$ denote the $n \times q$ design matrix whose $i$th row is the vector $Z_i'$ in equation (6). Assuming normal distributions, Wang applied results of Cook and Weisberg (1983) and showed that the score statistic for the null hypothesis is

$$D^W = 0.5R'Z(Z'Z)^{-1}Z'R/\hat{\sigma}^4,$$

where $R = (r_1^2, \ldots, r_n^2)'$, $r_i$ is the residual for $Y_i$ from fitting the location model by ordinary least squares, and $\hat{\sigma}$ is the maximum likelihood estimator of $\sigma = \exp(\phi_0/2)$ assuming common variance for all the observations. If none of the factors in $Z$ have dispersion effects, Wang showed that $D^W$ is distributed asymptotically as $\chi_q^2$.

For the orthogonal designs considered, the statistic $D^W$ can be decomposed as a sum of $q$ orthogonal components, one for each effect in $Z$. Thus Wang proposed testing the null hypothesis that factor $j$ has no dispersion effect by comparing the test statistic

$$D_j^W = 0.5(Z_j'R)^2/\hat{\sigma}^4 Z_j'Z_j$$

to a $\chi_1^2$ reference distribution. For two-level designs, $D_j^W$ is

$$D_j^W = 0.5(SS(j+) - SS(j-))^2/(n\hat{\sigma}^4)$$

and uses the same decomposition of the residual sum of squares as the Box–Meyer statistic. Wang (2001) showed how the method could be generalized to orthogonal designs with factors at more than two levels.

Brenneman and Nair (2001) showed that a slightly modified version of Wang's statistic can be interpreted as a normalized estimate of a dispersion effect if the following linear model is adopted for the variance, instead of (6),

$$\sigma_i^2 = \phi_0 + Z_i'\phi + v_i.$$

An unbiased estimator of $\phi_j$ can be obtained as follows. First fit an expanded location model, as in Bergman and Hynén (1997). Then estimate $\phi_j$ by

$$\hat{\phi}_j = (SS(j+) - SS(j-))/m(j),$$

where $m(j)$ is the degrees of freedom given by Bergman and Hynén (1997). See Arvidsson et al. (2003) for further discussion of this estimator.

For the dyestuffs experiment, the Wang statistic also finds that factor $F$ has a dispersion effect. With the main effects of $A$ and $F$ in the location model, $D_F^W = 8.13$ ($p$-value $= 0.005$). The next strongest effects are the $ADEF$ interaction ($D_{ADEF}^W = 4.10$, $p$-value $= 0.043$) and the $CEF$ interaction ($D_{CEF}^W = 4.01$, $p$-value $= 0.045$).

## 5.6 McGrath and Lin's Parametric Method

McGrath and Lin (2001) developed a test statistic motivated by the need to contend with multiple dispersion effects in two-level orthogonal designs. Suppose that factors $j$ and $u$ are being considered as having potential dispersion effects. The McGrath–Lin statistic is constructed from residuals from an expanded location model, as in the Bergman and Hynén (1997) approach. McGrath and Lin fitted a model that includes all identified location effects, the interactions of all such effects, and the interactions of all these effects with $j$, $u$, and the $ju$ interaction. If the expanded location model exhausts all the degrees of freedom in the experiment, the McGrath–Lin statistic cannot be applied. If not, the location model will divide the runs of a $2^{k-p}$ design into distinct sets in which the levels of the (expanded) location factors are constant within each set. Denote these sets by $C_t$, $t = 1, \ldots, T$, and let $S_t^2$ be the sum of squared residuals for runs in $C_t$. As $j$, $u$, and $ju$ are all included in the expanded location model, each will be at one level only in each set $C_t$. The McGrath–Lin statistic for factor $j$ is

$$D_j^{ML} = \left[ \frac{\prod_{j(+)} S_t^2}{\prod_{j(-)} S_t^2} \right]^{2/T},$$

with corresponding expressions for $u$ and for $ju$.

The McGrath–Lin statistic is designed to work well if there is a log-linear dispersion model with nonzero effects for at least two of the three effects being tested. If $\phi_j = 0$ in the dispersion model, McGrath and Lin showed that $D_j^{ML}$ has approximately an $F(c, c)$ distribution, where

$$c = \frac{2[\Gamma(d/2 + 2/T)\Gamma(d/2 - 2/T)]^{T/2}}{[\Gamma(d/2 + 2/T)\Gamma(d/2 - 2/T)]^{T/2} - \Gamma^T(d/2)}$$

and $d = n/T - 1$.

The main advantage of the McGrath–Lin (2001) statistic is reduced confounding when multiple dispersion effects are present. There are also several drawbacks. The statistic is complicated to compute and may be applicable only to a subset of the main effects and interactions in the design. For example, McGrath and Lin (2001) presented a $2^{5-1}$ experiment with four active location effects. Eight potential

dispersion effects could not be tested because they would have resulted in a fully saturated location model. A further practical problem is the need to specify first three potential dispersion effects ($j$, $u$, and the $ju$ interaction) and not just a single effect. In the dyestuffs experiment, there are more than 600 such triples, so that examining all of them is a major task and may lead to results that are difficult to interpret. An alternative is to begin the analysis with another method that identifies the most likely dispersion effects and then to use the McGrath–Lin statistic on the interaction triples involving those identified effects.

For the dyestuffs experiment, we applied the McGrath–Lin method to all 15 pairs of main effects with factors $A$ and $F$ in the location model. Factor $F$ consistently had a strong dispersion effect, with $p$-values of 0.0001 to 0.004, depending on the second factor in the pair. Factor $A$ was also found to have a potential dispersion effect, with $p$-values of 0.028 to 0.16. The $p$-value for factor $A$ was less than 0.05 except when paired with factor $C$. So the analyst is left with a practical problem of deciding whether factor $A$ has a dispersion effect. The analysis provides conflicting evidence about factor $A$ and it is not clear which pairing(s) should dictate the decision.

## 5.7 McGrath and Lin's Nonparametric Method

For all the test statistics presented thus far, reference distributions were derived by assuming normally distributed data. McGrath and Lin (2002) developed a nonparametric method that eliminates the need for this assumption. Their method is based on an alternative representation of the Bergman–Hynén (1997) statistic that exploits the formula derived by Box and Meyer (1986) to express sums of squares in terms of estimated regression coefficients. To test for a dispersion effect of factor $j$, McGrath and Lin (2002) proposed using location models that include factor $j$ and also its interactions with all active location effects, as in Bergman and Hynén (1997). The standard Bergman–Hynén (1997) statistic can then be computed from the regression coefficients for all effects *not* in the expanded location model. McGrath and Lin (2002) created a nonparametric version of the statistic by replacing the estimated regression coefficient $\hat{\beta}(j)_i$ by its rank $R(j)_i$ among all the coefficients not in the location model. The numerator and denominator in the rank version of the Bergman–Hynén (1997) statistic have a constant sum, so the test statistic can be constructed from the denominator alone,

$$D_j^{ML-NP} = \sum (R(j)_i - R(j)_u)^2,$$

with the sum extending over all pairs of effects for which $i \circ u = j$. McGrath and Lin (2002) used simulations to tabulate the distribution of $D_k^{ML-NP}$ for a small number of summands. They showed that a normal approximation could be applied when the number of summands is large.

The McGrath–Lin nonparametric statistic finds a number of dispersion effects in the dyestuffs experiment. The strongest effects are the main effect of $F$ and the $CEF$ interaction, which have $p$-values of 0.003 and 0.001, respectively. In addition, the main effect of $B$ and the $BC$, $BF$, $BDF$, and $ABCF$ interactions all

have $p$-values between 0.02 and 0.03. The number of terms in the McGrath–Lin statistics is large for this experiment, so the $p$-values were computed using the normal approximation.

## 5.8 Combined Location and Dispersion Modeling

Once tentative models for location and dispersion have been identified, they can be estimated by maximum likelihood. Most authors have advocated use of a log-linear model for dispersion effects. The estimation typically requires an iterative scheme. First the location effects are estimated and the residuals from the location model are used to estimate initial dispersion effects. Then the location effects are re-estimated by weighted least squares, with the weights computed from the estimated dispersion effects. These two steps are then iterated until a convergence criterion is satisfied. This type of approach has been described in the context of robust design experiments by Nelder and Lee (1991, 1998), Engel and Huele (1996), and Pan and Taam (2002). Lee and Nelder (1998) showed how to use restricted maximum likelihood for estimating the dispersion model. Wolfinger and Tobias (1998) applied ideas developed in the context of mixed linear models to a more general setting that includes location, dispersion, and random effects. For more general presentations of joint modeling of location and dispersion, see Aitkin (1987), Carroll and Ruppert (1988), and Verbyla (1993). A linear model for the dispersion effects has also been considered by some authors (Brenneman and Nair, 2001; Arvidsson et al., 2003). One limitation to the use of the above models is that they are not available as standard options in most software packages. Thus some custom programming is needed to fit them.

There is consensus among those who have written about dispersion effects that combined location and dispersion models can provide good estimates of all effects. The major source of controversy is about how to identify the models and, in particular, the dispersion model.

We analyzed the dyestuffs data using a log-linear main effects model for the dispersion. The dispersion effect of $F$ is significant with a coefficient of 0.54 and a standard error of 0.19 and the effect of $B$ is barely significant with a coefficient of $-0.40$ and a standard error of 0.19. If the $CEF$ interaction is added to the dispersion model, it proves to have a significant effect with a coefficient of 0.43 and a standard error of 0.19. We wonder, though, whether many scientists would be prepared to adopt a model with a three-factor interaction affecting the variance.

## 5.9 Brenneman and Nair's Method

Brenneman and Nair (2001) proposed a strategy that combines their modified version of Harvey's method with joint location and dispersion modeling for a log-linear dispersion model. After fitting a location model with ordinary least squares regression, they recommended an initial check to see if there are sufficient degrees of freedom even to consider looking for dispersion effects. The condition they

proposed is to look at the largest "closed" model contained in the location model, where a model is closed if all interactions among its active effects are also included in the model. If the largest closed model has at least $n/2$ terms, then the experiment is deemed to lack sufficient information to consider dispersion effects. If not, then they recommended using the modified Harvey method for initial identification. The next step is to fit a joint location–dispersion model, estimating the dispersion effects by restricted maximum likelihood. They cautioned that the dispersion model should include all effects found at the previous stage and all effects that correspond to interactions of two such effects. The reason for this is the potential for bias in the modified Harvey statistic for interactions of active dispersion effects. Some effects in the dispersion model may prove to be inert, so subsequent models deleting these effects might also be estimated by the same joint analysis.

When applied to the dyestuffs experiment, the Brenneman–Nair method determines that there are enough residual degrees of freedom to study dispersion effects. However, the method gets stuck at the next step due to the inability of the modified Harvey's method to handle the machine-zero residuals.

## *5.10 Other Methods*

In the interest of completeness, we briefly mention some additional methods that have been proposed.

Chowdhury and Fard (2001) presented a method for estimating dispersion effects from robust design experiments with right censored data. Kim and Lin (2002) proposed a method to determine optimal design factor settings that take account of both location and dispersion effects when there are multiple responses. They based their approach on response surface models for location and dispersion of each response variable.

Liao (2000) derived a test statistic for single dispersion effects in $2^{n-k}$ designs. He applied the generalized likelihood ratio test for a normal model to the residuals after fitting a location model, which results in Bartlett's (1937) classical test for comparing variances in one-way layouts. The test is then applied, in turn, to compare the variances at the two levels of each of the $k$ experimental factors. We caution that the test statistic (equation (3) in Liao) is written incorrectly.

Holm and Wiklander (1999) presented test statistics for dispersion effects derived from quadratic functions of the location effects. These test statistics are equivalent to the Box and Meyer (1986) statistics. The Holm and Wiklander version emphasizes how they can be seen as correlation coefficients among the null location effects, which can be used as a basis for making statistical inferences about possible dispersion effects.

## *5.11 Location–Dispersion Confounding*

Correct identification of the location model is a serious problem that affects all the methods for identifying and estimating dispersion effects. Pan (1999) showed that small to moderate location effects that are undetected can seriously impair

subsequent identification of dispersion effects in the methods proposed by Bergman and Hynén (1997) and Box and Meyer (1986). The other methods described here are also affected by missed location effects. Pan found this to be a sufficiently serious problem that he argued against the use of unreplicated designs for identifying dispersion effects.

Pan's results on confounding of location and dispersion effects can be understood from our equations (4) and (5) (Box and Meyer, 1986) relating sums of squared residuals to the effects *dropped* from the location model. In most experiments, some location effects will be not quite large enough to warrant inclusion in the location model. Take the two largest effects (in absolute value) that were dropped from the location model. The interaction of these effects is then likely to have an extreme Box–Meyer statistic $D_j^{BM}$ (with larger variance at the high level if the location effects have the same sign and at the low level if they have opposite signs).

McGrath and Lin (2002) suggested examining the contribution of each pair of coefficients to $D_j^{BM}$ as an effective diagnostic check. For example, one can then detect if just one pair of borderline location effects is responsible for a putative dispersion effect.

One might try to correct for the problems caused by moderate location effects by just fitting a larger location model. However, expanding the location model may leave too few degrees of freedom in the residuals to enable identification of dispersion effects.

The dyestuffs example illustrates the sensitivity of the dispersion analyses to the fitted location model. All the analyses reported thus far adjust for location effects of factors $A$ and $F$. However, the conclusions on dispersion effects are altered if the location model is expanded to include also the main effects of $B$ and $C$ and the $AC$ interaction. The strongest Box–Meyer statistic is now the $BEF$ interaction ($-0.42$), closely followed by the main effect of $F$ ($-0.40$). The main effect for $B$ is also of similar magnitude (0.36). For this location model, a normal plot of the Box–Meyer statistics does not show any effects that stand out from noise. The Bergman–Hynén and Wang statistics also find that the strongest effect is for the $BEF$ interaction (BH $= 2.57$, $p$-value $= 0.01$, Wang $= 4.26$, $p$-value $= 0.04$) followed by the main effect for $F$ (BH $= 2.49$, $p$-value $= 0.01$, Wang $= 3.92$, $p$-value $= 0.05$). The problem with Harvey's method is exacerbated because additional observations now have machine-zero residuals and they cause several effects to have similar dispersion statistics. The McGrath–Lin statistic also runs into problems because of machine-zero residuals. We again ran the method for all pairs of main effects and discovered unusual results when factor $E$ was one of the factors involved. Closer investigation revealed that one of the $T$ sets of residuals in these analyses had a machine-zero residual variance. Among the other factor pairs, only factor $F$ shows some possibility of having a dispersion effect. However, the $p$-value for factor $F$ is highly dependent on the factor with which it is paired. When paired with factor $A$, $B$, or $C$, it has a $p$-value of 0.013 (note that these analyses are identical, as all involve the same location model and therefore the same sets of residuals), but when paired with factor $D$, the $p$-value is 0.632. So the analyst is left with a quandary as to whether factor $F$ does, or does not, affect dispersion.

## 6 The Dyestuffs Example—Fractional Factorials

In Section 5, we introduced the dyestuffs experiment to illustrate the methods for screening for dispersion effects in unreplicated fractional factorial experiments. Typically we anticipate that smaller experiments will be used for screening. So, in this section, we analyze two sets of 16 runs that are extracted from the dyestuffs experiment and which constitute fractional factorials more typical of the actual size of screening experiments.

The first fraction has generators $E = ABC$ and $F = ABD$ and the second fraction has generators $E = BCD$ and $F = ACD$. In both fractions the dominant location effects are the main effects for factors $A$ and $F$, exactly as in the full factorial. In the second fraction, these effects clearly stand out from all the rest and are identified as significant by Lenth's method. In the first fraction, $A$ and $F$ are not as well distinguished from the next largest effects. Lenth's method finds a significant effect for $A$, but $F$ just barely falls below the initial cutoff for significance. Given the concern that missed location effects can lead to erroneous conclusions about dispersion effects, we would recommend including $A$ and $F$ in the location model for both fractions.

The Box–Meyer, Bergman–Hynén, and Wang statistics, for both fractions, point to factor $F$ as having the most potential for a dispersion effect, with weaker evidence for the contrast associated with both the $CD$ and $EF$ interactions. In the first fraction, the Box–Meyer statistic for $F$ is 0.81, compared to $-0.63$ for the $CD = EF$ interaction. The Bergman–Hynén statistic for $F$ is almost statistically significant with a $p$-value of 0.064. The next strongest effects are the $CD = EF$ interaction and $E$, with $p$-values of 0.11 and 0.15, respectively. The Wang statistic for $F$ is 3.49 ($p$-value $= 0.13$) and that for $CD$ is 1.62 ($p$-value $= 0.20$).

The potential importance for dispersion of main effects for $E$ and $F$ and an $EF$ interaction suggests that this would be a good test case for the McGrath–Lin statistic. The appropriate location model includes main effects and all interactions of factors $A$, $E$, and $F$. That model has zero residuals for all four observations with $E$ and $F$ at their low values. (As in the full factorial, our software actually computes these as machine-zero.) Thus, the McGrath–Lin statistic will indicate strong effects for all three terms. However, the presence of the zero residuals casts questions on the validity of any distributional results for the statistic.

Use of restricted maximum likelihood to fit a main effects log-linear dispersion model, with only $A$ and $F$ in the location model, also suggests possible dispersion effects for $E$ and $F$, both of which have absolute effects slightly greater than 0.6. However, the standard errors of these coefficients are about 0.44, so they fall well short of being statistically significant. Adding the $EF$ interaction to the dispersion model leads to highly significant effects for all three terms, with coefficients of 2.46 for $E$, 2.89 for $F$, and $-2.91$ for $EF$. Here, the location model includes only $A$ and $F$ and does not have any zero residuals. Nonetheless, the residual sums of squares arising from a simple linear regression on the location model differ substantially across the combinations of $E$ and $F$ (see Table 2).

TABLE 2. Sums of squares of the residuals (SSq) for levels of $E$ and $F$ for the dyestuffs experiment, with only $A$ and $F$ in the location model.

| $E$ | $F$ | SSq in $2^6$ | SSq in first fraction |
|---|---|---|---|
| −1 | −1 | 202 | 23 |
| −1 | 1 | 670 | 396 |
| 1 | −1 | 356 | 78 |
| 1 | 1 | 592 | 113 |

Do $E$ and $EF$ have dispersion effects, as indicated from this fraction? It is informative to return to the full experiment to examine this question. The residual sums of squares from a simple linear regression on $A$ and $F$ are shown in Table 2. A dispersion model with all three effects has a coefficient of 0.47 for $F$, similar to that found earlier, but the coefficients for $E$ and $EF$ are 0.06 and −0.16, respectively. Thus the full experiment provides no evidence at all of dispersion effects for $E$ or $EF$.

In the second fraction, the Box–Meyer, Bergman–Hynén, and Wang statistics indicate that $F$ is the only factor that affects dispersion. The Box–Meyer statistic for $F$ is 1.14, compared to 0.62 for the next strongest contrast, which is associated with both the $BC$ and $DE$ interactions. Another contrast, associated with only three-factor interactions, is of similar strength. The Bergman–Hynén statistic for $F$ is clearly significant with a $p$-value of 0.006. The same two contrasts with extreme values of the Box–Meyer statistic have $p$-values of about 0.11 each. The Wang statistic for $F$ is 3.49 ($p$-value = 0.062). The same contrasts as above are next in line, but with $p$-values above 0.20. We also computed the McGrath–Lin statistic for possible joint dispersion effects of $E$, $F$, and $EF$. As with the first fraction, all the residuals were zero when both $E$ and $F$ were at low levels, suggesting effects for all three contrasts but leaving questions as to statistical inference for the effects.

Combined location–dispersion models also find significant dispersion effects for factors $D$ and $E$. With $A$ and $F$ in the location model and a log-linear dispersion model, the effects of $D$, $E$, and $F$ are 1.34, 1.75, and 2.84, respectively, all with approximate standard errors of 0.4 to 0.45. As with the first fraction, we identify dispersion effects that are not present in the full data set.

The modified Harvey's method proves problematic in both fractions because of zero residuals.

## 7 Discussion

Identifying factors that affect dispersion and estimating their effects can be of great value. The understanding of dispersion effects can be used to improve the quality of manufactured products. Outside the quality domain, knowledge of dispersion effects can be crucial for statistical modeling and as input into the design of experiments or surveys. We have reviewed here the numerous methods that

have been presented in the last 15 years to study dispersion effects with screening designs.

Our review has led us to two main conclusions.

- The best way to screen for dispersion effects is to include noise factors in the experiment and to exploit noise factor by control factor interactions. Experiments such as this will be most successful when the noise factors are indeed responsible for a large share of the outcome variation. Taguchi's idea of using noise factors to force controlled variation into experimental data is a striking and important contribution.
- The identification of dispersion effects from unreplicated experiments is a risky business. If there is a single factor with a dominant dispersion effect, a number of methods (such as those of Box and Meyer, 1986; Bergman and Hynén, 1997; Wang 1989; or modified Harvey) appear to be reasonably successful. But, in more complex settings, there is considerable risk of wrongly identifying effects as present when they are not, or of missing effects that really are present. As we have shown with our examples, the test statistics and more sophisticated methods such as joint location–dispersion modeling can both go awry in the small unreplicated experiments that are often used for screening location effects. We can only issue the standard caution of *caveat emptor*.

There are several reasons why unreplicated experiments pose great difficulty for identifying dispersion effects. The first problem is the requirement of specifying the location effects. As shown by Pan (1999) (following on the derivations of Box and Meyer, 1986), statistics for dispersion effects can depend critically on the choice of location model. A second problem is the possible bias in the dispersion statistics when there is more than one dispersion effect. The methods proposed to date to overcome these problems do not appear to be adequate. The modified Harvey's method (Brenneman and Nair, 2001) could not be applied to our example due to zero residuals. We point out that zero residuals will be quite unusual in the simulation studies used to explore these methods, because simulated data have high resolution. But in real experiments, the resolution of observations is limited and we have seen many experiments with some zero residuals. The same problem caused difficulty for the McGrath–Lin (2001) statistic. For the latter statistic, there is the additional problem of specifying in advance which interaction triples should be studied. The practitioner is faced with the circular problem of needing to identify the dispersion effect triples in order to apply the identification procedure. Finally, there is the problem observed clearly in our examples that results from a small experiment may not reflect the patterns found when more data are collected.

Given results from a small, unreplicated screening experiment, we would recommend the following procedure.

1. If the location model is too large, do not attempt to study dispersion effects. The Brenneman–Nair (2001) condition is a good guideline.
2. Use the Bergman–Hynén (1997) method as a quick screen for dominant dispersion effects.

3. To check for possible additional dispersion effects, fit joint location–dispersion models. These methods are easier to implement than the statistics proposed for identifying multiple dispersion effects. They give both estimates of the strength of the dispersion effects and approximate test statistics.
4. Regard the results with due caution! If at all possible, collect some additional data to verify any conclusions.

## *References*


Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics*, **36**, 332–339.

Ankenman, B. E. and Dean, A. M. (2001). Quality improvement and robustness via design of experiments. In *Handbook of Statistics*, Volume 22, Chapter 8. Editors: R. Khattree and C. R. Rao. Elsevier, Amsterdam.

Arvidsson, M., Bergman, B., and Hynén, A. (2003). Comments on dispersion effect analysis. Unpublished manuscript.

Arvidsson, M., Merlind, P. K., Hynén, A., and Bergman, B. (2001). Identification of factors influencing dispersion in split-plot experiments. *Journal of Applied Statistics*, **28**, 269–283.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *Supplement to the Journal of the Royal Statistical Society*, **4**, 137–183.

Bergman, B. and Hynén, A. (1997). Dispersion effects from unreplicated designs in the $2^{k-p}$ series. *Technometrics*, **39**, 191–198.

Bérubé, J. and Nair, V. N. (1998). Exploiting the inherent structure in robust parameter design experiments. *Statistica Sinica*, **8**, 43–66.

Blomkvist, O., Hynén, A., and Bergman, B. (1997). A method to identify dispersion effects from unreplicated multilevel experiments. *Quality and Reliability Engineering International*, **13**, 127–138.

Borkowski, J. J. and Lucas, J. M. (1997). Designs of mixed resolution for process robustness studies. *Technometrics*, **39**, 63–70.

Box, G. E. P. (1988). Signal-to-noise ratios, performance criteria and transformations. *Technometrics*, **30**, 1–40.

Box, G. E. P. and Draper, N. R. (1987). Empirical Model-Building and Response Surfaces. John Wiley and Sons, New York.

Box, G. E. P. and Meyer, R. D. (1986). Dispersion effects from factorial designs. *Technometrics*, **28**, 19–27.

Brenneman, W. A. (2000). Inference for location and dispersion effects in unreplicated factorial experiments. PhD Dissertation, University of Michigan, Ann Arbor.

Brenneman, W. A. and Nair, V. N. (2001). Methods for identifying dispersion effects in unreplicated factorial experiments. *Technometrics*, **43**, 388–405.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall, London.

Chowdhury, A. H. and Fard, N. S. (2001). Estimation of dispersion effects from robust design experiments with censored response data. *Quality and Reliability Engineering International*, **17**, 25–32.

Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, **70**, 1–10.

Engel, J. and Huele, A. F. (1996). A generalized linear modeling approach to robust design. *Technometrics*, **38**, 365–373.

Ferrer, A. J. and Romero, R. (1993). Small samples estimation of dispersion effects from unreplicated data. *Communications in Statistics: Simulation and Computation*, **22**, 975–995.

Ferrer, A. J. and Romero, R. (1995). A simple method to study dispersion effects from non-necessarily replicated data in industrial contexts. *Quality Engineering*, **7**, 747–755.

Hedayat, S. and Stufken, J. (1999). Compound orthogonal arrays. *Technometrics*, **41**, 57–61.

Holm, S. and Wiklander, K. (1999). Simultaneous estimation of location and dispersion in two-level fractional factorial designs. *Journal of Applied Statistics*, **26**, 235–242.

Kim, K. J. and Lin, D. K. J. (2006). Optimization of multiple responses considering both location and dispersion effects. European Journal of Operational Research, **169**, 133–145.

Lee, Y. and Nelder, J. A. (1998). Generalized linear models for analysis of quality-improvement experiments. *Canadian Journal of Statistics,* **26**, 95–105.

Lenth, R. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, **31**, 469–473.

Liao, C. T. (2000). Identification of dispersion effects from unreplicated $2^{n-k}$ fractional factorial designs. *Computational Statistics and Data Analysis*, **33**, 291–298.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. Chapman & Hall, London.

McGrath, R. N. and Lin, D. K. J. (2001). Testing multiple dispersion effects in unreplicated fractional factorial designs. *Technometrics*, **43**, 406–414.

McGrath, R. N. and Lin, D. K. J. (2002). A nonparametric dispersion test for unreplicated two-level fractional factorial designs. *Journal of Nonparametric Statistics*, **14**, 699–714.

Montgomery, D. C. (1990). Using fractional factorial designs for robust process development. *Quality Engineering*, **3**, 193–205.

Montgomery, D. C. (1999). Experimental design for product and process design and development. *Journal of the Royal Statistical Society D*, **38**, 159–177.

Myers, R. H. and Montgomery, D. C. (2002). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, second edition. John Wiley and Sons, New York.

Myers, R. H., Khuri, A. I., and Vining, G. (1992). Response surface alternatives to the Taguchi robust parameter design approach. *The American Statistician*, **46**, 131–139.

Myers, R. H., Montgomery, D. C., and Vining, G. G. (2002). *Generalized Linear Models with Applications in Engineering and the Sciences*. John Wiley and Sons, New York.

Nair, V. N. and Pregibon, D. (1988). Analyzing dispersion effects from replicated factorial experiments. *Technometrics*, **30**, 247–257.

Nelder, J. A. and Lee, Y. (1991). Generalized linear models for the analysis of Taguchi-type experiments. *Applied Stochastic Models and Data Analysis*, **7**, 107–120.

Nelder, J. A. and Lee, Y. (1998). Joint modeling of mean and dispersion. *Technometrics*, **40**, 168–175.

Pan, G. (1999). The impact of unidentified location effects on dispersion effects identification from unreplicated factorial designs. *Technometrics*, **41**, 313–326.

Pan, G. and Taam, W. (2002). On generalized linear model method for detecting dispersion effects in unreplicated factorial designs. *Journal of Statistical Computation and Simulation*, **72**, 431–450.

Phadke, M. S. (1989). Quality Engineering Using Robust Design. Prentice-Hall, Englewood Cliffs, NJ.

Rosenbaum, P. (1994). Dispersion effects from fractional factorials in Taguchi's method of quality design. *Journal of the Royal Statistical Society B*, **56**, 641–652.

Rosenbaum, P. (1996). Some useful compound dispersion experiments in quality design. *Technometrics*, **38**, 354–364.

Shoemaker, A. C., Tsui, K.-L., Wu, C. F. J. (1991). Economical experimentation methods for robust design. *Technometrics*, **33**, 415–427.

Steinberg, D. M. (1996). Robust design: Experiments for improving quality. In *Handbook of Statistics*, 13, Chapter 7, pages 199–240. Editors: S. Ghosh and C. R. Rao. Elsevier, Amsterdam.

Steinberg, D. M. and Bursztyn, D. (1994). Dispersion effects in robust design experiments with noise factors. *Journal of Quality Technology*, **26**, 12–20.

Steinberg, D. M. and Bursztyn, D. (1998). Noise factors, dispersion effects and robust design. *Statistica Sinica*, **8**, 67–85.

Taguchi, G. (1986). *Introduction to Quality Engineering*. Unipub/Kraus International, White Plains, New York.

Verbyla, A. P. (1993). Modelling variance heterogeneity: Residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society B*, **55**, 493–508.

Vining, G. G. and Myers, R. H. (1990). Combining Taguchi and response surface philosophies: A dual response approach. *Journal of Quality Technology*, **22**, 38–45.

Wang, P. C. (1989). Tests for dispersion effects from orthogonal arrays. *Computational Statistics and Data Analysis*, **8**, 109–117.

Wang, P. C. (2001). Testing dispersion effects from general unreplicated fractional factorial designs. *Quality and Reliability Engineering International*, **17**, 243–248.

Wolfinger, R. and Tobias, R. (1998). Joint estimation of location, dispersion and random effects in robust design. *Technometrics*, **40**, 62–71.

Wu, C. F. J. and Hamada, M. (2000). Experiments: Planning, Analysis and Parameter Design Optimization. John Wiley and Sons, New York.

# 3
# Pooling Experiments for Blood Screening and Drug Discovery

Jacqueline M. Hughes-Oliver

Pooling experiments date as far back as 1915 and were initially used in dilution studies for estimating the density of organisms in some medium. These early uses of pooling were necessitated by scientific and technical limitations. Today, pooling experiments are driven by the potential cost savings and precision gains that can result, and they are making a substantial impact on blood screening and drug discovery. A general review of pooling experiments is given here, with additional details and discussion of issues and methods for two important application areas, namely, blood testing and drug discovery. The blood testing application is very old, from 1943, yet is still used today, especially for HIV antibody screening. In contrast, the drug discovery application is relatively new, with early uses occurring in the period from the late 1980s to early 1990s. Statistical methods for this latter application are still actively being investigated and developed through both the pharmaceutical industries and academic research. The ability of pooling to investigate synergism offers exciting prospects for the discovery of combination therapies.

## 1 Introduction

The use of pooling experiments began as early as 1915 and was, initially, used in dilution studies for estimating the density of organisms in some medium. Examples quoted by Halvorson and Ziegler (1933) include investigations of densities of bacteria in milk and protozoa in soil. Prior to 1915, most dilution methods were inadequate because they failed to account for chance or error in observation. In 1915, McCrady presented a method of estimation based on probability which was then expanded by Halvorson and Ziegler (1933) to provide an estimator of density based on pooled data. Fisher (1921) also used a similar pooling-based estimator.

These early uses of pooling experiments were born of necessity, as explained below for bacterial density estimation. In order to determine the absence or presence of bacteria in a fluid, cultures are made of a number of samples (small amounts) of the fluid. Growth of a colony of bacteria within a fluid sample indicates the presence of bacteria and no growth indicates absence of bacteria. The act of culturing this fluid can be viewed as applying a test, the result of which is "good"

or "bad". This test is applied simultaneously to every molecule present in that sample of fluid and the results for all the molecules are pooled. The combined test results (from all samples) are then used to estimate the density of bacteria present in the source fluid. Although it is virtually impossible to perform this test on individual molecules, it is quite simple to ascertain whether a culture from the pooled molecules is free of colony growth.

Today, pooling studies are not typically used from necessity. Rather, they are used because of the economic gains, savings in time, or precision gains that can result. A useful review of pooling experiments from the point of view of composite sampling methods is offered by Lancaster and Keller-McNulty (1998). This chapter focuses on current usage for populations in which individuals are labeled with respect to one or more traits and where pooling experiments are optional. More specifically, the discussion addresses applications in blood testing and drug discovery. This is not meant to be an exhaustive review, but rather a vehicle for highlighting some important aspects of pooled screening in these two areas.

Applications in drug discovery require the identification of "hit compounds," which are those compounds having activity greater than some prespecified threshold in one or more biological assays. Good hit compounds need to be identified quickly to allow progression to other phases of drug discovery (see Chapter 4). One application in blood screening requires the identification of individuals with sero-prevalence (detectability in blood) of one or more diseases. Cost effectiveness is important here because a balance must be struck between the cost of testing, which can be high, and the large populations that must be screened. A second issue that arises in blood screening is the need to estimate prevalences, possibly as a function of covariates.

In order to address the two areas of application simultaneously, the term *individual* is used to mean either a person (in the context of blood testing) or a compound (drug discovery); the term *active* means either positive for one or more diseases (blood testing) or exceeding an activity threshold (drug discovery); and the term *population* means either a group of people being screened (blood testing) or a compound library being screened (drug discovery).

Two fundamentally different problems arise from pooling experiments, namely, *estimation* and *classification*. Estimation involves the use of pooled samples for decreasing the cost-per-unit information when estimating the prevalence of active individuals in a population. These estimation results may then be used as the end-product of analysis or they may be incorporated into a classification scheme. The estimation results serve as the end-product of analysis when the goal of the study is to estimate the prevalence of active individuals but there is no interest expressed in actually identifying these active individuals. In a classification scheme, the ultimate goal is screening for the purpose of identifying active individuals. The performance of a classification scheme is typically assessed by considering the expected number of tests required to identify active individuals with particular attributes. Drug discovery is considered to be a classification problem, but results from the estimation problem can also be used to inform classification decisions. For blood testing, the

application may be either a classification or an estimation problem, depending on the context.

The most common assumptions of pooling experiments are briefly critiqued in Section 2. In Section 3, general accomplishments and advances in pooling experiments are reviewed, irrespective of their particular applications. Sections 4 and 5 provide details specific to blood testing and drug discovery, respectively.

## 2 Types of Pooling and Assumptions

Pooling experiments are of two basic types, *simple* or *orthogonal*. In simple pooling, each individual appears in exactly one pool; see Figure 1(a), where each circle in the box represents an individual, and individuals in the same column are in the same pool. An active pool response must be followed by individual testing to determine which specific individuals in the pool are active. In orthogonal pooling

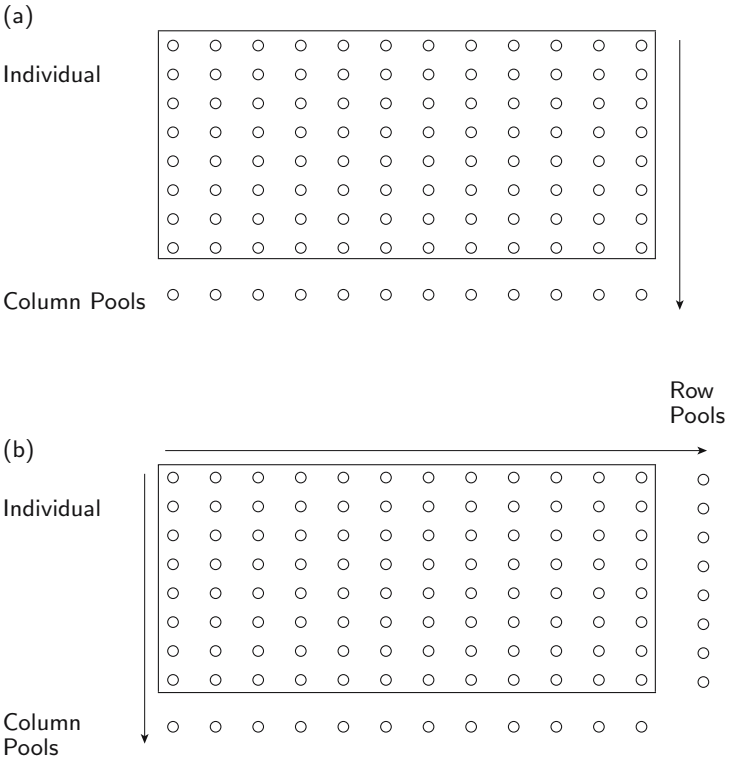



FIGURE 1. (a) Simple pooling: Individuals, represented by circles, are pooled according to their column location; each individual appears in exactly one pool. (b) Two-way orthogonal pooling: Individuals are pooled according to both their column and row locations; each individual appears in exactly two pools.

using $d$ dimensions, individuals appear in exactly $d$ pools; see Figure 1(b) for $d = 2$. The determination of active individuals from orthogonal pooling is easier than that from simple pooling. Consider orthogonal pooling over $d = 2$ dimensions corresponding to rows and columns. If a compound lies simultaneously in an active row and an active column, then it is reasonable to believe that this compound is active and that it should thus be assigned a favorable rank for individual testing. Despite its benefits, orthogonal pooling adds many complexities and is, consequently, not as popular as simple pooling. All further discussion is limited to simple pooling.

Pooling experiments are based, historically, on several assumptions that are often blatantly unjustified. The first assumption is that individuals have equal probabilities of being active. In blood testing, genetic characteristics, environmental exposures, and demographic identities are widely accepted as sources of variability for disease status, thus suggesting that probabilities of activity are not constant across the population (Dorfman, 1943). In drug discovery, it is well recognized that structure–activity relationships (SARs, see Chapter 4), where activity is related to chemical structural features of a compound, lead to nonconstant probabilities of activity; see McFarland and Gans (1986).

A second assumption generally used is that interactions do not occur within a pool; that is, activity is neither enhanced nor degenerated by testing multiple compounds using a single test on a pool. It is possible, however, that individually inactive compounds can give an active test result when pooled together (Borisy et al., 2003), thus providing a case of "activity enhancement" by pooling. This phenomenon is called *synergism* and its detection is crucial to the development of combination therapies in the pharmaceutical industry. The reverse situation can also occur in that pooled testing of individually diseased samples can result in disease-free pool results (Phatarfod and Sudbury, 1994), thus providing a case of "activity degeneration" by pooling. This phenomenon is called *antagonism* or *blocking* and is considered an undesirable potential effect of pooling in the blood testing application. Blocking relationships that occur in drug discovery applications can have a positive impact on screening outcomes in that they provide further implicit evidence of structure–activity relationships.

A third assumption concerns absence of errors in testing. Both blood testing and drug discovery have strong potential for false negatives and false positives. Errors in testing are inherently linked to assumptions regarding interactions within a pool. Both concepts are, in turn, related to the sometimes arbitrarily chosen threshold value used for categorizing a continuous assay response into only two classes of "active" or "inactive".

## 3 History of Pooling Experiments

Dorfman (1943) has been credited with the origin of pooling experiments in the statistical literature. His ideas were popularized through the books of Feller (1957, page 225) and Wilks (1962) and became known as "the blood testing problem".

Many efforts were then made to refine Dorfman's proposal by extending the number of stages using various retesting schemes and by relaxing assumptions. This section provides a brief summary of some key results from these efforts; see also Chapter 9 for related work in factorial experiments.

## *3.1 The Dorfman Model and Assumptions*

Suppose that, in a large population of $f$ individuals, each individual has, independently, the same probability $p$ of being active. In this context, $p$ represents a latent propensity for an individual to be active; some individuals ultimately express this latent feature and are thus labeled as actives, whereas others never express the latent feature and thus are labeled as inactive. If individuals are pooled into groups of size $k$ and if pooling does not alter the behavior of individuals, then the resulting $g = f/k$ pools will, independently, have the same probability $\theta = 1 - (1 - p)^k$ of being active. Hence, the number of active pools, $X$, follows a binomial distribution with parameters $g$ and $\theta$. Of course, activity of pools or individuals must be revealed by some testing system and for now this system is assumed to be perfect. In other words, *sensitivity* (the probability that a test will identify, by testing outcome, an individual as active given that the individual is truly active) and *specificity* (the probability that a test will identify, by testing outcome, an individual as inactive given that the individual is truly inactive) are both assumed to be 1.0. Dorfman himself did not believe these assumptions strictly but was able to build from the strength of the overall approach to make worthwhile reductions in the required number of tests over one-at-a-time testing. Aspects of sensitivity and specificity are also discussed in Chapters 4 and 6.

### 3.1.1 Classification

Dorfman's application was the need to identify World War II Selective Service inductees whose blood contained syphilitic antigens. In other words, his was a classification problem and he wanted to minimize the number of tests required to classify all inductees. All individuals in inactive pools were declared to be inactive, without further testing. All individuals in active pools were subjected to one-at-a-time testing, thus leading to a random total number of tests $T = f/k + Xk$ (where $f$, $k$, and $X$ are defined above). Dorfman then needed to determine a pool size to minimize the expected total number of tests. Pooling would only be advantageous if, on average, the total number of tests is less than $f$, which is the number of tests required by one-at-a-time testing. Dorfman minimized the expected relative cost, for given $p$,

$$\frac{E(T)}{f} = \frac{1}{k} + 1 - (1 - p)^k, \qquad \text{for } k > 1,$$

with respect to $k$, to determine the best possible improvements offered by pooling experiments over one-at-a-time testing. For example, by pooling, he obtained an 80% cost savings in tests over one-at-a-time testing when $p = .01$ and $k = 11$. The savings decrease as $p$ increases but are still appreciable even for larger $p$ with, for example, 28% savings when $p = .15$ and $k = 3$. In fact, pooling, based

on Dorfman retesting for classification, is better than one-at-a-time testing when $1/k < (1-p)^k$. The approximation $(1-p)^k \approx e^{-pk}$ is sometimes used to claim that pooling is better than one-at-a-time testing when $p < (\ln k)/k$. When $p$ and $k$ are both large, many active pools are observed and, consequently, more individual retests are required and this reduces the desirability of pooling.

Despite its simplicity, Dorfman's retesting strategy is still very widely used today, especially in blood testing and drug discovery applications. His rough guidelines for choosing $k$ such that $p < (\ln k)/k$, coupled with recommendations by Thompson (1962), Kerr (1971), Loyer (1983), and Swallow (1987) to use an a priori upper bound on $p$, is also commonly used today. Indeed, the attraction of the Dorfman strategy is its simplicity. Improved methods for classification, some of which are discussed in this article, add various levels of complications that users may not yet be ready to accept.

### 3.1.2 Estimation

Dorfman (1943) did not really address the problem of estimating the prevalence $p$ but, using his assumptions, others did. Gibbs and Gower (1960) and Thompson (1962) investigated the maximum likelihood estimator of $p$:

$$\hat{p} = 1 - \left(1 - \frac{X}{g}\right)^{1/k},$$

where $g = f/k$ is the number of pools and $X$ the number of active pools.

This is a positively biased, but consistent, estimator for $p$. Based on the asymptotic variance of $\hat{p}$, Peto (1953) and Kerr (1971) determined that the optimum group size $k$ satisfies $(1-p)^k = .203$. Based on asymptotic considerations, Thompson (1962) suggested that the group size should be approximately $k = (1.5936/p) - 1$. He also argued, however, that the asymptotic results can be very misleading and offered small-sample exact bias and variance formulae. Gibbs and Gower (1960), Griffiths (1972), Loyer (1983), and Swallow (1985, 1987) also gave small-sample results.

When $c$ is the nontesting cost associated with obtaining an individual sample (for example, personnel time for drawing blood from an individual) divided by the cost of a test, Sobel and Elashoff (1975) showed that pooling is advantageous when

$$p < 1 - \frac{1+2c}{3+2c}.$$

For extremely costly tests, pooling can be beneficial for $p$ as large as $2/3$.

## *3.2 Some Alternative Models*

Extensions of Dorfman's procedure follow four main branches:

(i) Development of different retesting schemes;
(ii) Strategies when $p$ is unknown, as is usually the case;
(iii) Departures from binomial assumptions; and
(iv) Errors in testing.

The literature is quite extensive (see Hughes-Oliver, 1991), so only key papers are referenced here. For brevity, no attempt has been made to separate extensions for the goal of classification from extensions for the goal of estimation.

### 3.2.1 Retesting Schemes

Many different retesting schemes have been suggested in the literature, some of which require infinite testability of the units. For example, Sterrett (1957) proposed retesting individuals in an active pool only until an active individual is found. The remaining untested individuals are then retested as a single pool and the process is repeated. Sobel and Groll (1959) proposed a retesting scheme based on *nested halving* procedures. Active pools are subdivided into two pools of size approximately $k/2$, each of which is tested. Individuals in an inactive subpool are declared inactive but an active subpool is again halved. Halving terminates when pool size becomes one, that is, at individual testing. Sobel and Elashoff (1975) proposed a general nested retesting scheme for estimation, of which nested halving is a special case. They found that a certain class of nested halving procedures is highly efficient and the savings over one-at-a-time procedures is even greater for the estimation problem than for the classification problem. They also found that, when the cost of obtaining individuals relative to the cost of a test is negligible, the optimal testing scheme does not include retesting. Chen and Swallow (1990) confirmed the finding that retesting is not advantageous for estimation when testing costs far exceed costs of obtaining individuals, but they showed that data from retesting can provide useful information for testing model assumptions.

In contrast to the work of Sobel and Elashoff (1975) and Chen and Swallow (1990), where the stated goal was to reduce cost per unit information for estimation in the presence of perfect testing, retesting has been shown to be useful for classification, especially when test results may be inaccurate. Litvak et al. (1994) argued that, even when testing is correctly executed, it can lead to incorrect conclusions and, in these cases, retesting provides significant improvements over no-retesting for reducing error rates associated with labeling samples when screening low-risk HIV populations.

Based on nested halving, Litvak et al. (1994) also proposed a new retesting scheme where inactive pools are subjected to a repeat test; if they again test inactive then all individuals in those pools are declared inactive, otherwise the pool is halved and subjected to additional testing. Gastwirth and Johnson (1994), who were also concerned with error rates for labeling individuals assuming imperfect testing, proposed a "back-end" retesting stage where pooled testing is used to rescreen a subset of individuals who were declared inactive from "first-stage" pooled testing.

### 3.2.2 Choosing the Pool Size

The success of a pooling experiment depends heavily on the choice of a good value for the pool size $k$. Unfortunately, optimal pool size depends on the value of $p$. In the absence of a priori information on $p$, Le (1981) and a number of other authors

recommended that different pool sizes be used and the resulting data on the number of active pools for each pool size be combined to yield an estimator. Thompson (1962) argued that an a priori upper bound on $p$ should be used to determine a single pool size, and Hughes-Oliver and Swallow (1994) and Hughes-Oliver and Rosenberger (2000) proposed two-stage adaptation to allow a single update of the pool size. These last authors also addressed the issue of pool size when there are multivariate responses from pools, motivated by the need to monitor prevalence rates for several diseases simultaneously.

### 3.2.3 Departures from Binomial Assumptions

On the issue of departures from binomial assumptions, Finucan (1964) considered a case where stratification occurs and results in different probabilities of activity for different individuals. A good early reference for various approaches to dealing with such situations is that of Hwang (1984). Chen and Swallow (1990) noted that model assumptions can be tested if data on unequal pool sizes are available. Many recent articles also consider the situation where probability of activity is dependent on covariates. For small numbers of covariates, Hung and Swallow (2000), Vansteelandt et al. (2000), Xie (2001), and Tebbs and Swallow (2003a,b) obtained estimates of prevalences in the different strata. For large numbers of covariates, Xie et al. (2001), Zhu et al. (2001), and Yi (2002) obtained estimates of prevalences in the different strata then ranked the estimated prevalences to define a testing order for the classification problem. Thus, the estimation problem was an intermediate step, not the ultimate goal, of the drug discovery applications of these authors. On a related note, Remlinger et al. (2005) considered the design problem of assigning individuals to pools based on their covariates; the goal was classification in the presence of covariate-dependent prevalences.

### 3.2.4 Errors in Testing

The problem of errors in testing has been examined by a host of investigators. References to investigators from a clinical/laboratory science viewpoint are given in Section 4.2. From a statistician's viewpoint, Gastwirth and Hammick (1989) and Hammick and Gastwirth (1994) used trinomial models in which either a confirmatory pool test or an independent pool test was used to reduce the number of false positives. They also incorporated sensitivities and specificities (Section 3.1) of the testing scheme into their estimator while maintaining individual anonymity. Tu et al. (1994, 1995) also incorporated sensitivities and specificities of the testing scheme and showed that this leads to improved estimation accuracy. Vansteelandt et al. (2000) took the same approach but with the added complication of covariate-adjusted estimation of prevalence. Hung and Swallow (1999) investigated robustness properties of the pooling estimator with respect to dilution effects and serial correlation models. Wein and Zenios (1996) also investigated dilution effects. In the area of drug discovery, Langfeldt et al. (1997), Xie et al. (2001), Zhu et al. (2001), Yi (2002), and Remlinger et al. (2005) all investigated

procedures that model possible interactions occurring within pools; such interactions may be mislabeled as errors in testing.

# 4 Pooling for Blood Screening

## *4.1 Background*

Pooling is now considered to be a routine option in blood screening, especially for the human immunodeficiency virus (HIV). There are many reports espousing the benefits of pooled testing in countries across the world, using a variety of assay techniques.

There are actually three blood screening applications for which pooling has been beneficial. The two most common applications are the context of classification, where the goal of blood screening is to identify individuals with sero-prevalence of one or more diseases. One classification application arises from the need to screen donated blood and blood products and the other from the need to screen for individual diagnoses. Cost effectiveness, as measured by the reduction in the expected total number of tests, is the most commonly used assessment of pooling methods. The third application is the need to monitor changes in sero-prevalence over time for (possibly) different sets of individuals, where demarcation of individuals may occur along demographic lines or spatial/regional clusters.

### 4.1.1 Classification

Motivated by a more than 90% transmission rate of HIV by transfusion of blood and blood products, the World Health Organization (WHO) argued for 100% screening of donated blood. Recognizing that developing countries can ill-afford the cost of 100% one-at-a-time screening, WHO issued recommendations for testing for HIV antibody on serum pools (WHO, 1991) in areas where sero-prevalence is less than 2%. In fact, this figure of 2% sero-prevalence is much too restrictive. Many investigators have achieved success with much higher prevalences. For example, Soroka et al. (2003) described the successful use of pooling where prevalence was 9%. It is important to note that, for screening blood supplies, complete identification of sero-positive individuals is not necessary. All that is needed is a method for tracking the complete donated sample, without personal identifiers. This makes pooled screening very attractive for screening blood supplies because donors can be assured that their anonymity will be maintained.

Another classification problem occurs when individual diagnosis is the required outcome of a screening campaign. In such a campaign, personal identifiers must be maintained for the purpose of reporting back to individuals about their sero-prevalence. Moreover, diagnostic testing requires that sero-positive pools be subjected to confirmatory gold-standard tests.

### 4.1.2 Estimation

Gastwirth and Hammick (1989) and Hammick and Gastwirth (1994) approached the blood testing problem with a keen eye towards preserving individual privacy rights. They proposed screening strategies designed for estimating prevalences. Rather than focus on the cost-saving advantages of pooling, these authors selected pooling because of the anonymity it provides to individuals being screened. They also reduced false predictive values by employing confirmatory tests to verify sero-prevalence.

## *4.2 HIV Testing*

The standard practice in developed countries for determining HIV sero-prevalence is first to apply the cost-effective, but suboptimal, enzyme-linked immunosorbent assay (ELISA) test. For those individuals who are identified as sero-positive by the ELISA test, follow-up testing is then performed using the gold standard Western blot test. Unfortunately, the Western blot is very expensive, difficult to standardize and often results in no clear diagnosis for some individuals (Tamashiro et al., 1993). To relieve the cost burden, the WHO recommends a series of repeat testing that uses cheaper tests, namely ELISA or simple or rapid tests, to avoid the Western blot while still maintaining testing accuracy. In general, the Western blot best is up to six times as expensive as rapid or simple tests and 18 times as expensive as ELISA; see, for example, WHO (1992). Rapid and simple tests provide results in less than one hour (less than 30 minutes for rapid tests) and may be performed by personnel having little or no laboratory training. ELISA must be performed in a laboratory (so results are not immediately available) by extensively trained laboratory professionals.

The WHO (1992) recommendations are shown in Figure 2 and supporting text is given in Table 1. Strategy I is recommended for screening contributions to a blood supply. It says that a contribution should only be accepted if it is sero-negative according to either the ELISA test, or the rapid test, or the simple test. Sero-positive samples are not considered further. Strategy I is also recommended when prevalence is high and the goal is HIV surveillance.

WHO's Strategy III is recommended for diagnosing symptom-free individuals living in areas of low prevalence. It is the strategy that allows the greatest number of retests. If the first test is sero-positive, it is followed up with a second test that is not simply a repeat measurement of the first test. Specifically, the assay procedure should differ from the first assay procedure in some substantial way; for example, different antigen preparation, different test principle (such as indirect versus competitive) or both. The first test should be very sensitive but the other two tests should have higher specificity than the first. If this second test is again sero-positive, a third and last test is applied. Strategy II is similar but with only two stages.

Effective clinical pooling studies for HIV classification and surveillance have been reported by a large number of investigators. Emmanuel et al. (1988),

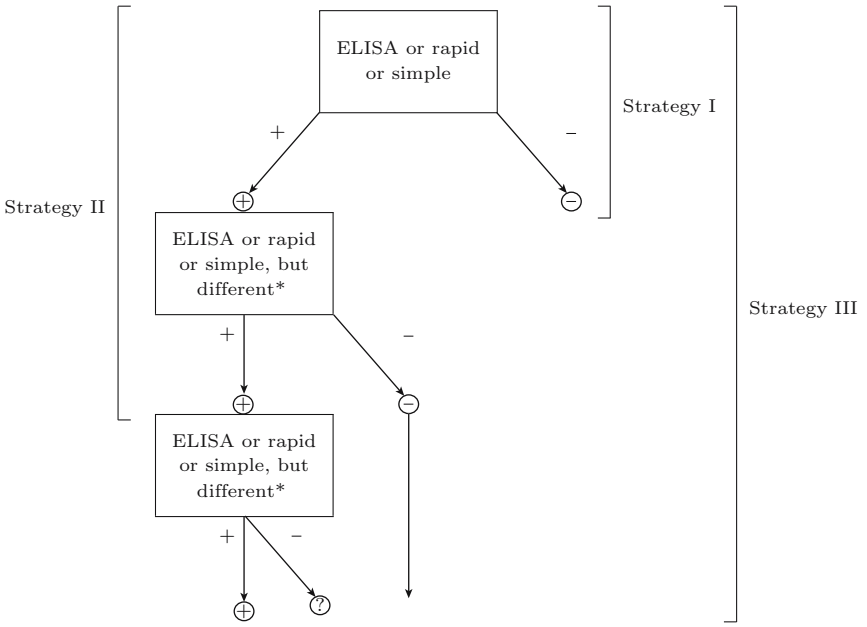



FIGURE 2. The World Health Organization 1992 recommendations on how to screen for HIV without using the Western blot test. * denotes that subsequent assays should differ from previous assays.

Cahoon-Young et al. (1989), Kline et al. (1989), Behets et al. (1990), Archbold et al. (1991), Ko et al. (1992), Babu et al. (1993), and Perriens et al. (1993) have all reported successes for several different countries, including countries in Africa and Asia. "Success" here is defined as the appropriate management of the logistics of pooling and the reduction of the amount of testing required. Moreover, successes have been achieved based on several different testing protocols, including ELISA, Western blot, and rapid testing techniques; see also Davey et al. (1991), Seymour et al. (1992), Raboud et al. (1993), McMahon et al. (1995), Verstraeten et al. (1998), and Soroka et al. (2003). These studies reported up to 80% reductions in cost for pooling experiments compared with one-at-a-time testing.

TABLE 1. WHO recommendations for the use of the strategies shown in Figure 2

| Strategy | Limits on $p$ | Objective |
|---|---|---|
| I | None | Blood supply |
| I | $p > 0.1$ | HIV surveillance |
| II | $p \leq 0.1$ | HIV surveillance |
| II | None | Diagnosis, HIV symptoms |
| II | $p > 0.1$ | Diagnosis, no symptoms |
| III | $p \leq 0.1$ | Diagnosis, no symptoms |

Since the late 1980s, statistical contributions to pooling for blood testing have focused on the following aspects: assessing changes in sensitivity and specificity due to pooling, designing pooling strategies to accommodate both cheap initial screens and gold-standard confirmatory screens, and estimation of covariate-dependent prevalences.

Let us first consider approaches to assessing changes in sensitivity and specificity due to pooling. As defined in Section 3.1, sensitivity is the probability that a test correctly detects antibodies in a serum sample, and specificity is the probability that a test correctly identifies an antibody-free serum sample. These probabilities have been a major area of concern in pooling studies for blood testing (WHO, 1991). The over-arching issue when screening a blood supply is whether dilution effects will cause a single sero-positive individual to be missed when combined in a pool with several (perhaps as many as 14) sero-negative individuals. This issue relates to the false negative predictive value as follows. A predictive value is the probability of truth given an individual's testing outcome; a false negative predictive value is the probability that the individual is truly active but is labeled as inactive from testing; a false-positive predictive value is the probability that the individual is truly inactive but is labeled as active from testing. When screening for diagnostic purposes, the major concern is that sero-negative individuals will be labeled sero-positive; this relates to the false positive predictive value. Repeatedly, however, studies have indicated that, under their worst performance, these possible pooling effects are negligible. In fact, Cahoon-Young et al. (1989), Behets et al. (1990), Archbold et al. (1991), Sanchez et al. (1991) all reported reductions in the number of misclassified sero-negative individuals; for example, Cahoon-Young et al. (1989) found that there were seven misclassified sero-negative individuals out of 5000 tested, but no misclassified sero-negative pools out of 500 tested.

For understanding sensitivity, specificity, false negative predictive value, and false positive predictive value, consider the four cells and two column margins of Table 2, where individuals are cross-classified with respect to their true sero-status versus observed sero-status. Sensitivity is represented by $S_e = P(\text{testing outcome } + \mid \text{truth is } +)$ and specificity is $S_p = P(\text{testing outcome } - \mid \text{truth is } -)$. With these definitions and with $p$ denoting the probability of an individual having

TABLE 2. Cross-classification of individuals for "true" versus "observed" sero-status $(+, -)$ in terms of sensitivity $S_e$, specificity $S_p$, and probability $p$ of positive sero-status

| | | True: + | True: − |
|---|---|---|---|
| Observed | + | $pS_e$ | $(1-p)(1-S_p)$ |
| | − | $p(1-S_e)$ | $(1-p)S_p$ |

positive sero-status, the false negative predictive value is

$$FNPV = P(\text{ truth is } + \mid \text{testing outcome } -) = \frac{p(1-S_e)}{p(1-S_e)+(1-p)S_p}$$

and the false positive predictive value is

$$FPPV = P(\text{ truth is } - \mid \text{testing outcome } +) = \frac{(1-p)(1-S_p)}{(1-p)(1-S_p)+pS_e}.$$

Large false negative predictive values are particularly troubling when screening a blood supply because they allow sero-positive samples to enter the blood supply system, thus leading to possible transmission of deadly diseases. Minimizing the false negative predictive value is probably more important than increasing cost efficiency of pooling for this application. Of course, large false negative predictive values can arise even when screening is accomplished using one-at-a-time testing. False positive predictive values are of greater concern in diagnostic testing because they can cause undue stress for the falsely identified individuals and increase testing costs. Notice that if $S_e = S_p = 1$, then $FNPV = FPPV = 0$ and no misclassifications will occur.

Litvak et al. (1994) compared three pooling strategies and one-at-a-time testing with respect to their abilities to reduce *FNPV*, *FPPV*, the expected numbers of tests required, and the expected numbers of tests performed for each individual. The first pooling study considered was Dorfman retesting with pool size $k = 15$; that is, all individuals in sero-positive pools were tested one-at-a-time but no retesting was applied to individuals in sero-negative pools. The pool size of 15 was selected because, at the time, it was the largest acceptable size from a laboratory perspective for maintaining high sensitivity and specificity after pooling. Litvak et al. (1994) called this screening protocol $T_0$. Their second pooling protocol, $T_2$, was essentially the retesting method proposed by Sobel and Groll (1959) whereby sero-positive pools are recursively halved and testing of the subpools continues until no further splits are possible. In this strategy with $k = 15$, a serum sample must be positive four or five times before being declared sero-positive. Their third pooling protocol, $T_2^+$, is similar to $T_2$ except that each sero-negative pool is subjected to one confirmatory pool test before all its individuals are labeled as sero-negative. It was found that $T_2$ and $T_2^+$ were comparable and that both provided huge reductions in *FPPV* compared with one-at-a-time testing but smaller reductions compared with $T_0$. For *FNPV*, $T_2^+$ was the best protocol. In short, pooling reduced both false negative and false positive predictive values.

The result from estimating sero-prevalence of HIV in the presence of errors in testing is really quite startling. Tu et al. (1994, 1995) found that pooling actually increases estimator efficiency by reducing the effect of measurement errors. Vansteelandt et al. (2000) extended the procedure to account for covariate adjustments. These results, along with the large number of empirical findings from investigators such as Emmanuel et al. (1988), clear the way for heavy reliance on pooling strategies to eliminate the backlog and reduce the cost of screening large populations. This is of particular importance to developing countries that are often

cash-strapped but might benefit the most from 100% screening. Even developed countries might want to rethink their screening strategies to take advantage of fewer but more informative pooled test results.

# 5 Pooling for Screening in Drug Discovery

## *5.1 Background*

Twenty percent of sales from the pharmaceutical industry for the year 2000 were reinvested into research and development activities. This percentage is higher than in most other industries, including the electronics industry. At the same time, it is getting increasingly difficult to introduce (that is, discover, demonstrate efficacy and safety, and receive approval for marketing of) new drugs in order to recoup investment costs. On average, one new drug requires investment of $880 million and 15 years development (Giersiefen et al., 2003, pages 1–2). The days of profitability of "runner-up" or "me-too" drugs have long passed and the simple current reality is that survival and financial security of a pharmaceutical company demands that they find the best drugs as fast as possible. This means that the five major phases of drug discovery, as illustrated in Figure 3, need to be traversed aggressively. Details on the phases of drug discovery can be found in Chapter 4. Here, attention is directed to the third phase, Lead Identification, which is where pooling experiments for screening in drug discovery usually occur.

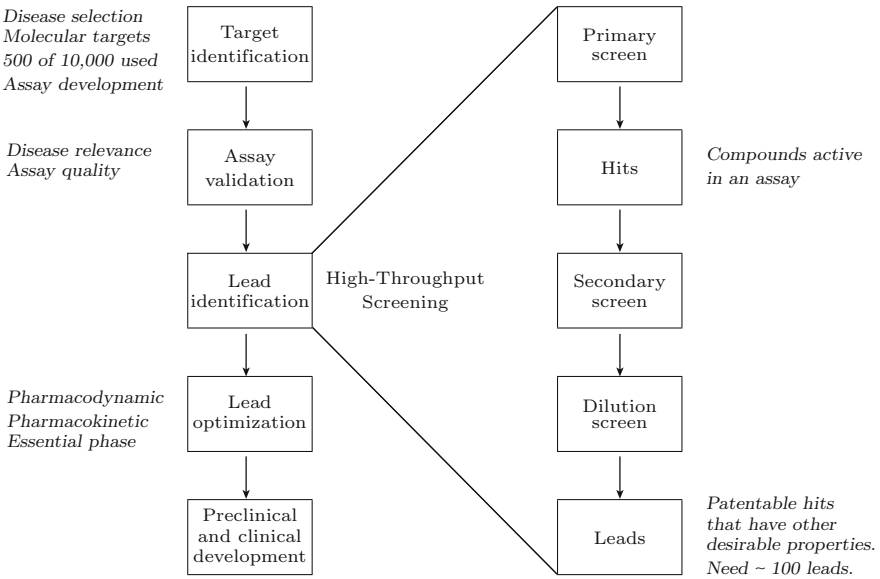



FIGURE 3. Phases of drug discovery.

Given a large collection of compounds, say $f = 500{,}000$, the goal of lead identification is to find about 100 compounds that are

(i) Active for the assay—this allows them to be called "hits";
(ii) Patentable; that is, their structures are novel and not already under patent;
(iii) Have good chemical properties such as stability, can be synthesized, are not toxic, and so on;
(iv) Something is understood about what makes them active; that is, their structure–activity relationships have been, at least partially, identified;
(v) Each compound is fairly different from the other ninety-nine.

Compounds that satisfy all these requirements are called *leads* or *lead compounds*. The need for properties (i)–(iii) is clear, but additional comments are warranted for the other properties.

Knowledge of structure–activity relationships allows chemists to focus on the essential substructures of the compound without wasting time with the portions that do not affect activity. The drug discovery phase that follows lead identification is *lead optimization*. In this phase, chemists expend enormous energies "tweaking" the leads to increase the chances of compounds making it through the grueling stages of preclinical and clinical development. It is imperative that the lead optimization phase produces very strong lead compounds to be investigated during preclinical and clinical development. Once a compound reaches the preclinical and clinical development phase, extensive additional financial and time investments are made, so that heavy losses would be incurred if the compound had to be abandoned further down the drug discovery channel because it possesses undesirable features (see also Chapter 4).

## *5.2 Differences and Similarities Between Blood Screening and Screening for Drug Discovery*

The goals of drug discovery, as stated above, seem to be very similar to those of the blood screening for classification problem, but this is not at all the case. As mentioned, in earlier sections of this chapter, approaches to solving the blood testing for classification problem do not routinely incorporate covariate information. For the HIV blood testing problem, relevant covariate information for an individual may include the following: number of blood transfusions received, number of sexual partners, number of sexual partners who are HIV-infected, syringe use, drug use, sexual preference, and HIV status of parents. Recent investigations have allowed the estimation of prevalence in different covariate-defined strata, but the number of strata is never large and is quite typically less than 10. In screening for drug discovery, on the other hand, the number of covariates is quite often at least twice the number of pooled responses available. Indeed, the significant challenges that arise from the high-dimensional-with-low-sample-size data sets that usually result from "high-throughput screening" in drug discovery present major obstacles to analysis, even for one-at-a-time testing results. These difficulties are magnified

in the presence of pooled responses. More information is given by Langfeldt et al. (1997), Xie et al. (2001), Zhu et al. (2001), Yi (2002), and Remlinger et al. (2005).

Arguably, the biggest difference between the two application areas discussed in this chapter is the potential for synergistic relationships between compounds in pools for drug discovery, whereas no such concept has arisen for blood testing. Synergism has recently become the major supporting argument for pursuing pooling experiments in drug discovery (Xie et al., 2001; Yi, 2002; Remlinger et al., 2005). Synergistic relationships can only be discovered through pooling studies where compounds are forced together, and it is these synergistic relationships that form the basis of *combination therapies*. These therapies involve deliberate mixing of drugs and they are now the standard of care for life-threatening diseases such as cancer and HIV. Current combination therapies were discovered by combining individually active compounds after they had been approved by the Food and Drug Administration. By investigating synergistic relationships in vitro, it is expected that one could find a combination where, individually, the compounds are inactive but, when pooled, their activities exceed all other combinations. Borisy et al. (2003) demonstrated this quite nicely using several real experiments. For example, chlorpromazine and pentamidin were more effective than paclitaxel (a clinically used anticancer drug), even though individually neither drug was effective at tolerable doses. Similar ideas were discussed by Tan et al. (2003).

So, are cost considerations no longer important for drug discovery? The answer is "not really," or at least not as much as they used to be. Before the advent of high-throughput screening (HTS, see Chapter 4) and ultrahigh-throughput screening (uHTS), pooling was necessary for processing the large compound libraries typically encountered. In those days, a large screening campaign might screen a total of 50,000 compounds, and it would take months to complete. Today, uHTS can screen 100,000 compounds in a single day; see Banks (2000) and Niles and Coassin (2002). HTS and uHTS systems are centralized, highly automated, and are under robotic control so they can work almost around the clock with very small percentages of down-time.

The two applications of drug discovery and blood testing are similar in how they process screening outcomes. Comparing Strategy III of Figure 2 with the extended view of Lead Identification in Figure 3, it can be seen that both methods use three tests in labeling the final selected individuals. The selected individuals are the gems for drug discovery applications but, for the blood testing problem, they actually cause concern because they are blood samples that have been confirmed to be diseased.

## *5.3 Design and Analysis Techniques*

A commonly used technique for analyzing drug discovery screening data from individuals is recursive partitioning (RP), more commonly known as "trees" (see, for example, Blower et al., 2002). In very recent times, efforts based on multiple trees (Svetnik et al., 2003) have become the method of choice, despite the additional difficulties associated with them, because of their good predictive abilities.

The number of researchers working to develop methodology appropriate for pooled drug screening data and who are allowed to discuss these issues outside the big pharmaceutical companies is very small. Papers from these researchers have been reviewed earlier in this chapter, but a few additional comments are warranted. The bulk of the work has been divided into two major paths. One path concerns the search for the efficient placement of individuals within pools; that is, the design of pooling studies. Because of the very large number of covariates, this is a difficult problem that requires computer-intensive techniques. Remlinger et al. (2005) obtained structure-based pooling designs to assign pool placement in response to covariate-adjusted prevalences. Zhu (2000) developed model-based designs for the same problem.

The second major path concerns analysis methods, including nonparametric, semi-parametric, fully parametric, and Bayesian approaches. Nonparameteric results are based on recursive partitioning on pooled data and require the formation of pooled summaries and decisions of whether and how to include the retested data in the analysis without violating independence assumptions. For the semi-parametric work, Yi (2002) modeled data from pooling experiments as missing data scenarios where missingness occurs at random. This was a novel use of the semi-parametric methodology to an area that had never before been considered. Another interesting finding is that random retesting of both active and inactive pools can lead to improved estimators. Litvak et al. (1994) and Gastwirth and Johnson (1994) were able to improve their estimators in the blood testing problem by retesting inactive pools.

Zhu et al. (2001) described a trinomial modeling approach that incorporates the phenomenon of blocking and used this model to develop criteria for creating pooling designs. These fully parametric models were also extended by Yi (2002) who considered pairwise blocking probabilities. Xie et al. (2001) used a Bayesian approach for modeling blockers and synergism. Finally, Remlinger et al. (2005) also considered design pooling strategies, but from a completely structure-based approach.

When it comes to designing and analyzing pooling studies for drug discovery, many open questions remain. Single-dose pooling studies, which is an area still in its infancy, have been the focus of this chapter. Multiple-dose pooling studies, which constitute a more mature area of research and application, can bring yet another level of interesting questions and evidence of the utility of pooling; see, for example, Berenbaum (1989).

## 6 Discussion

Many modern developments and applications point to a bright future for pooling experiments. First, blood testing is ready to support heavy-duty use of pooling studies all across the world. The evidence of success is overwhelming whereas the costs are minimal. Secondly, the drug discovery application still has a long way to go before it is fully developed, but researchers are making great strides. The

ability to uncover synergistic relationships for discovering combination therapies is very exciting and offers many new challenges and possibilities.

## *References*


Archbold, E., Mitchel, S., Hanson, D., Hull, B., Galo, L. M., and Monzon, O. (1991). Serum-pooling strategies for HIV screening: Experiences in Trinidad, Ecuador, and the Phillippines. *VII International Conference on AIDS Abstract Book*, **2**, 331.

Babu, P. G., Saraswathi, N. K., Vaidyanathan, H., and John, T. J. (1993). Reduction of the cost of testing for antibody to human immunodeficiency virus, without losing sensitivity, by pooling sera. *Indian Journal of Medical Research Section A–Infectious Diseases*, **97**, 1–3.

Banks, M. (2000). Automation and technology for HTS in drug development. In *Approaches to High Throughput Toxicity Screening*. Editors: C. K. Atterwill, P. Goldfarb, and W. Purcell, pages 9–29. Taylor and Francis, London.

Behets, F., Bertozzi, S., Kasali, M., Kashamuka, M., Atikala, L., Brown, C., Ryder, R. W., and Quinn, T. C. (1990). Successful use of pooled sera to determine HIV-1 seroprevalence in Zaire with development of cost efficiency models. AIDS, **4**, 737–741.

Berenbaum, M. C. (1989). What is synergy? *Pharmacological Reviews*, **41**, 93–141.

Blower, P., Fligner, M., Verducci J., and Bjoraker, J. (2002). On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *Journal of Chemical Information and Computer Sciences*, **42**, 393–404.

Borisy, A. A., Elliott, P. J., Hurst, N. W., Lee, M. S., Lehar, J., Price, E. R., Serbedzija, G., Zimmerman, G. R., Foley, M. A., Stockwell, B. R., and Keith, C. T. (2003). Systematic discovery of multicomponent therapies. *Proceedings of the National Academy of Sciences*, **100**, 7977–7982.

Cahoon-Young, B., Chandler, A., Livermore, T., Gaudino, J., and Benjamin, R. (1989). Sensitivity and specificity of pooled versus individual sera in a human immunodeficiency virus antibody prevalence study. *Journal of Clinical Microbiology*, **27**, 1893–1895.

Chen, C. L. and Swallow, W. H. (1990). Using group testing to estimate a proportion, and to test the binomial model. *Biometrics*, **46**, 1035–1046.

Davey, R. J., Jett, B. W., and Alter, H. J. (1991). Pooling of blood donor sera prior to testing with rapid/simple HIV test kits. *Transfusion*, **31**, 7.

Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Science*, **14**, 436–440.

Emmanuel J. C., Bassett M. T., Smith H. J., and Jacobs, J. A. (1988). Pooling of sera for human immunodeficiency virus (HIV) testing: An economical method for use in developing countries. *Journal of Clinical Pathology*, **41**, 582–585.

Feller, W. (1957). *An Introduction to Probability Theory and its Applications*. John Wiley and Sons, New York.

Finucan, H. M. (1964). The blood testing problem. *Applied Statistics*, **13**, 43–50.

Fisher, R. A. (1921). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, *A*, **222**, 309–368.

Gastwirth, J. L. and Hammick, P. A. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of AIDS antibodies in blood donors. *Journal of Statistical Planning and Inference*, **22**, 15–27.

Gastwirth, J. L. and Johnson, W. O. (1994). A two-stage adaptive group-tesing procedure for estimating small proportions. *Journal of the American Statistical Association*, **89**, 972–981.

Gibbs, A. J. and Gower, J. C. (1960). The use of a multiple-transfer method in plant virus transmission studies: Some statistical points arising in the analysis of results. *Annals of Applied Biology*, **48**, 75–83.

Giersiefen, H., Hilgenfeld, R., and Hillisch, A. (2003). Modern methods of drug discovery: An introduction. In *Modern Methods of Drug Discovery*. Editors: A. Hillisch and R. Hilgenfeld. Birkhauser Verlag, Basel and Boston, pages 1–30.

Griffiths, D. A. (1972). A further note on the probability of disease transmission. *Biometrics*, **28**, 1133–1139.

Halvorson, H. O. and Ziegler, N. R. (1933). Application of statistics to problems in bacteriology. *Journal of Bacteriology*, **25**, 101–121.

Hammick, P. A. and Gastwirth, J. L. (1994). Group testing for sensitive characteristics: Extension to higher prevalence levels. *International Statistical Review*, **62**, 319–331.

Hann, M., Hudson, B., Lewell, X., Lifely, R, Miller, L., and Rarnsden, N. (1999). Strategic pooling of compounds for high-throughput screening. *Journal of Chemical Information and Computer Sciences*, **39**, 897–902.

Hughes-Oliver, J. M. (1991). Estimation using group-testing procedures: Adaptive iteration. PhD dissertation. North Carolina State University, Department of Statistics.

Hughes-Oliver, J. M. and Rosenberger, W. F. (2000). Efficient estimation of the prevalence of multiple rare traits. *Biometrika*, **87**, 315–327.

Hughes-Oliver, J. M. and Swallow, W. H. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association*, **89**, 982–993.

Hung, M. and Swallow, W. (1999). Robustness of group testing in the estimation of proportions. *Biometrics*, **55**, 231–237.

Hung, M. and Swallow, W. (2000). Use of binomial group testing in tests of hypotheses for classification or quantitative covariables. *Biometrics*, **56**, 204–212.

Hwang, F. K. (1984). Robust group testing. *Journal of Quality Technology*, **16**, 189–195.

Kerr, J. D. (1971). The probability of disease transmission. *Biometrics*, **27**, 219–222.

Kline, R. L., Brothers, T. A., Brookmeyer, R., Zeger, S., and Quinn, T. C. (1989). Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera. *Journal of Clinical Microbiology*, **27**, 1449–1452.

Ko, Y. C., Lan, S. J., Chiang, T. A., Yen, Y. Y., and Hsieh, C. C. (1992). Successful use of pooled sera to estimate HIV antibody seroprevalence and eliminate all positive cases. *Asia Pacific Journal of Public Health*, **6**, 146–149.

Lancaster, V. A. and Keller-McNulty, S. (1998). A review of composite sampling methods. *Journal of the American Statistical Association*, **93**, 1216–1230.

Langfeldt, S. A., Hughes-Oliver, J. M., Ghosh, S., and Young, S. S. (1997). Optimal group testing in the presence of blockers. *Institute of Statistics Mimeograph Series 2297*. North Carolina State University, Department of Statistics.

Le, C. T. (1981). A new estimator for infection rates using pools of variable size. *American Journal of Epidemiology*, **14**, 132–136.

Litvak, E., Tu, X. M., and Pagano, M. (1994). Screening for presence of a disease by pooling sera samples. *Journal of the American Statistical Association*, **89**, 424–434.

Loyer, M. (1983). Bad probability, good statistics, and group testing for binomial estimation. *The American Statistician*, **37**, 57–59.

McCrady, M. H. (1915). The numerical interpretation of fermentation-tube results. *Journal of Infectious Diseases*, **17**, 183–212.

McFarland, J. W. and Gans, D. J. (1986). On the significance of clusters in the graphical display of structure-activity data. *Journal of Medicinal Chemistry*, **29**, 505–514.

McMahon, E. J., Fang, C., Layug, L., and Sandler, S. G. (1995). Pooling blood donor samples to reduce the cost of HIV-1 antibody testing. *Vox Sanguinis*, **68**, 215–219.

Niles, W. D. and Coassin, P. J. (2002). Miniaturization technologies for high-throughput biology. *In Integrated Drug Discovery Technologies*. Editors: H.-Y. Mei and A. W. Czarnik, pages 341–364. Marcel Dekker, New York.

Perriens, J. H., Magazani, K., Kapila, N., Konde, M., Selemani, U. Piot, P., and van der Groen, G. (1993). Use of a rapid test and an ELISA for HIV antibody screening of pooled serum samples in Lubumbashi, Zaire. *Journal of Virological Methods*, **41**, 213–221.

Peto, S. (1953), A dose response equation for the invasion of micro-organisms. *Biometrics*, **9**, 320–335.

Phatarfod, R. M. and Sudbury, A. (1994). The use of a square array scheme in blood testing. *Statistics in Medicine*, **13**, 2337–2343.

Raboud, J. M., Sherlock, C., Schecter, M. T., Lepine D. G., and O'Shaughnessy, M. V. (1993). Combining pooling and alternative algorithms in seroprevalence studies. *Journal of Clinical Microbiology*, **31**, 2298–2302.

Remlinger, K. S., Hughes-Oliver, J. M., Young, S. S., and Lam, R. L. H. (2005). Statistical design of pools using optimal coverage and minimal collision. *Technometrics*, in press.

Sanchez, M., Leoro, G., and Archbold, E. (1991). Workload and cost-effectiveness analysis of a pooling method for HIV screening. *VII International Conference on AIDS Abstract Book*, volume 2, page 330.

Seymour, E., Barriga, G., and Stramer, S. L. (1992). Use of both pooled saliva and pooled serum samples for the determination of HIV-1/HIV-2 antibodies by both conventional and rapid EIA techniques. *International Conference on AIDS*, **8**, 194.

Sobel, M. and Elashoff, R. M. (1975). Group testing with a new goal, estimation. *Biometrika*, **62**, 181–193.

Sobel, M. and Groll, P. A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. *The Bell System Technical Journal*, **38**, 1179–1252.

Soroka, S. D., Granade, T. C., Phillips, S., and Parekh, B. (2003). The use of simple, rapid tests to detect antibodies to human immunodeficiency virus types 1 and 2 in pooled serum specimens. *Journal of Clinical Virology*, **27**, 90–96.

Sterrett, A. (1957). On the detection of defective members of large populations. *Annals of Mathematical Statistics*, **28**, 1033–1036.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., and Feuston, B.P. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, **43**, 1947–1958.

Swallow, W. H. (1985). Group testing for estimating infection rates and probability of disease transmission. *Phytopathology*, **75**, 882–889.

Swallow, W. H. (1987). Relative mean squared error and cost considerations in choosing group size for group testing to estimate infection rates and probabilities of disease transmission. *Phytopathology*, **77**, 1376–1381.

Tamashiro, H., Maskill, W., Emmanuel, J., Fauquex, A., Sato, P., and Heymann, D. (1993). Reducing the cost of HIV antibody testing. *Lancet*, **342**(8863), 87–90.

Tan, M., Fang, H. B., Tian, G. L., and Houghton, P. J. (2003). Experimenal design and sample size determination for testing synergism in drug combination studies based on uniform measures. *Statistics in Medicine*, **22**, 2091–2100.

Tebbs, J. M. and Swallow, W. H. (2003a). Estimating ordered binomial proportions with the use of group testing. *Biometrika*, **90**, 471–477.

Tebbs, J. M. and Swallow, W. H. (2003b). More powerful likelihood ratio tests for isotonic binomial proportions. *Biometrical Journal*, **45**, 618–630.

Thompson, K. H. (1962). Estimation of the proportion of vectors in a natural population of insects. *Biometrics*, **18**, 568–578.

Tu, X. M., Litvak, E., and Pagano, M. (1994). Screening tests: Can we get more by doing less? *Statistics in Medicine*, **13**, 1905–1919.

Tu, X. M., Litvak, E., and Pagano, M. (1995). On the information and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika*, **82**, 287–297.

Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*, **56**, 1126–1133.

Verstraeten, T., Farah, B., Duchateau, L., and Matu, R. (1998). Pooling sera to reduce the cost of HIV surveillance: A feasibility study in a rural Kenyan district. *Tropical Medicine and International Health*, **3**, 747–750.

Wein, L. and Zenios, S. (1996). Pooled testing for HIV screening: Capturing the dilution effect. *Operations Research*, **44**, 543–569.

World Health Organization (WHO). (1991). Recommendations for testing for HIV antibody on serum pool. *World Health Organization Weekly Epidemiological Record*, **66**(44), 326–327.

World Health Organization (WHO) (1992). Recommendations for the selection and use of HIV antibody tests. *World Health Organization Weekly Epidemiological Record*, **67**(20), 145–149.

Wilks, S. S. (1962). *Mathematical Statistics*. John Wiley and Sons, New York.

Xie, M. (2001). Regression analysis of group testing samples. *Statistics in Medicine*, **20**, 1957–1969.

Xie, M., Tatsuoka, K., Sacks, J., and Young, S. S. (2001). Group testing with blockers and synergism. *Journal of the American Statistical Association*, **96**, 92–102.

Yi, B. (2002). Nonparametric, parametric and semiparametric models for screening and decoding pools of chemical compounds. Unpublished PhD dissertation. North Carolina State University, Department of Statistics.

Zhu, L. (2000). Statistical decoding and designing of pooling experiments based on chemical structure. Unpublished PhD dissertation. North Carolina State University, Department of Statistics.

Zhu, L., Hughes-Oliver, J. M., and Young, S. S. (2001). Statistical decoding of potent pools based on chemical structure. *Biometrics*, **57**, 922–930.

# 4
# Pharmaceutical Drug Discovery: Designing the Blockbuster Drug

David Jesse Cummins

Twenty years ago, drug discovery was a somewhat plodding and scholastic endeavor; those days are gone. The intellectual challenges are greater than ever but the pace has changed. Although there are greater opportunities for therapeutic targets than ever before, the costs and risks are great and the increasingly competitive environment makes the pace of pharmaceutical drug hunting range from exciting to overwhelming. These changes are catalyzed by major changes to drug discovery processes through application of rapid parallel synthesis of large chemical libraries and high-throughput screening. These techniques result in huge volumes of data for use in decision making. Besides the size and complex nature of biological and chemical data sets and the many sources of data "noise", the needs of business produce many, often conflicting, decision criteria and constraints such as time, cost, and patent caveats. The drive is still to find potent and selective molecules but, in recent years, key aspects of drug discovery are being shifted to earlier in the process. Discovery scientists are now concerned with building molecules that have good stability but also reasonable properties of absorption into the bloodstream, distribution and binding to tissues, metabolism and excretion, low toxicity, and reasonable cost of production. These requirements result in a high-dimensional decision problem with conflicting criteria and limited resources. An overview of the broad range of issues and activities involved in pharmaceutical screening is given along with references for further reading.

## 1 Introduction

The pharmaceutical industry is rapidly approaching a crisis situation. Epidemics, such as AIDS, increasing rates of cancer, the threat of biological warfare agents, and an increasing elderly population, mean that the demand for useful therapeutic drugs is greater than ever. At the same time, pressure is increasing to reduce costs in the face of the daunting challenge of discovering and developing therapeutic agents. Approximately 50% of drugs in development fail due to safety issues and 25% fail due to efficacy issues. Most researchers estimate that the process of developing a new drug from early screening to drug store shelves costs $600 to $900 million and takes 8 to 15 years.

In this chapter, a *screen* refers to a biochemical test (or *assay*) to see if small molecules bind to a target. The usual sense of this term suggests an experiment

performed on some (physical) experimental unit. There is a hierarchy of results: human clinical trials are the ultimate answer, which are approximated with animal testing, animal testing is approximated with in vitro testing (cell cultures, enzyme studies), and any of the above can be approximated *in silico* by the use of predictive models (a *virtual* screen).

Although the greatest expenses in drug discovery and development are incurred in the clinical trials phases, this chapter focuses on the early screening stage, before the first human dose. Well-planned studies at this stage have great potential to reduce expenses at later stages of the process. If it were possible to weed out the toxic molecules prior to the clinical trials phase, fully 40% of the expenses incurred in clinical trials would be eliminated! Even a small dent in this expensive process would result in enormous savings. If this could be done through virtual screens using predictive models, then additional savings would be achieved through less animal toxicity testing and this would also reduce the overall drug development time.

Drug discovery is a multidisciplinary endeavor with critical work at the interface of biology, chemistry, computer science, and informatics. In biology, a major activity is to make the linkages between what can be assayed and a disease response, but activities also include design and validation of animal models, cell cultures, biochemical screen design, and assay variability studies. Another important area in biology, the pace of which has especially intensified in the last decade, is the assessment in vivo of the extent of absorption into the blood stream, distribution and binding to tissues, and the rates of metabolism and excretion. This is denoted by *ADME* and is discussed in Section 11. In chemistry, the major responsibility is to provide the creative spark to navigate effectively the large space of possible compounds (another word for molecules) towards the blockbuster drug. Other activities include synthesis of new molecules, analytical characterization of existing molecules (including purity of batches, pKa, logP, melting point, and solubility) and construction of libraries. Important issues in computer science include data storage and extraction, implementation and scale-up of algorithms, management of biological and chemical databases, and software support. Activities in informatics or chemoinformatics (Leach, 2003) include design of experiments, development of new chemical descriptors, simulation, statistical analysis, mathematical modeling, molecular modeling, and the development of machine learning algorithms.

The successful development of a new drug depends on a number of criteria. Most importantly, the drug should show a substantial beneficial effect in the treatment of a particular disease (*efficacy*). This implies that, in addition to intrinsic activity, the drug is able to reach its *target* (a biological gateway that is linked to a disease state—a large organic molecule that may be a receptor, a protein, or an enzyme) and does not produce overwhelming toxic effects. Many active drugs fail in later phases of the development process because they do not reach their intended target.

The main challenges in drug discovery fall into four categories:

1. Potency: the drug must have the desired effect, in the desired time frame, at a low dosage.

2. Selectivity: the drug should produce only the desired activity and not cause side effects. There are so many possible targets in the body that achieving high selectivity is difficult. Side effects may be caused by metabolites of the drug, by-products produced when the body uses enzymes to break down the drug in the elimination process (Section 11.1).
3. ADME and pharmacokinetics (or *PK* ): the drug must reach the site of action. If taken orally, it must be water soluble, survive the stomach, be absorbed through the intestine, survive attack by many enzymes, and be transported into the target cells across the cell membrane. It must not be metabolised too quickly, but also must not be so stable or protein bound that it accumulates in the body. Another important factor to control is whether a compound crosses the blood–brain barrier (or BBB).
4. Toxicity: there are many mechanisms by which a compound can be toxic. Toxicity issues may arise from PK or ADME or selectivity issues, depending on the mechanism. Alternatively a compound may simply react harmfully with tissues or organs in a direct manner.

One may think of an iterative model for the preclinical discovery screening cycle. A large number of compounds are to be mined for compounds that are *active*; for example, that bind to a particular target. The compounds may come from different sources such as vendor catalogues, corporate collections, or combinatorial chemistry projects. In fact, the compounds need only to exist in a virtual sense, because in silico predictions in the form of a model can be made in a *virtual screen* (Section 8) which can then be used to decide which compounds should be physically made and tested. A mapping from the structure space of compounds to the *descriptor space* or *property space* provides covariates or explanatory variables that can be used to build predictive models. These models can help in the selection process, where a subset of available molecules is chosen for the biological screen. The experimental results of the biological screen (actives and inactives, or numeric potency values) are then used to learn more about the *structure–activity relationship* (*SAR*) which leads to new models and a new selection of compounds as the cycle renews.

The relationship between the biological responses and the changes in chemical structural motifs is called SAR or QSAR (*quantitative structure–activity relationship*). Small changes to the chemical structure can often produce dramatic changes in the biological response; when this happens, chemists and biologists will often describe the SAR as *nonlinear*, by which they mean that the SAR has a "sensitive" or "rough" or "unstable" response surface. Often the chemical compounds are considered to be the experimental unit even though, in actual experiments, an animal or cell culture is the unit and the compound is a treatment. This is because the important asset to the pharmaceutical company is the compound. The vast size of the set of potential experimental units (potential compounds), coupled with the high dimensionality of the response being optimized (potency, selectivity, toxicity, and ADME) and the "roughness" of the response landscape make drug discovery

a challenging arena. The level of noise in biological data can be extremely high as well.

This chapter covers a selection of problems and case study examples. Perspectives specific to Eli Lilly and Company (or Lilly) are distinguished from broader perspectives believed to be shared by most companies across the industry. The chapter covers both design and analysis issues, and touches on topics such as simulation, computer experiments, and pooling. Section 2 gives an overview of drug design. In Section 3 the issue of false negatives and false positives in drug screening is addressed. Molecular diversity is discussed in Section 4, and machine learning is the topic of Section 6. Section 7 describes a lower-throughput iterative approach to screening and virtual screening in drug discovery projects in an iterative *Active Learning* strategy. A brief mention of pooling strategies is made in Section 9 and Section 10 discusses expectations for rare events. Section 11 describes aspects of a molecule that determine its ability to be safely transported to the area of the body where it can be of therapeutic benefit. Finally, in Section 12 the problem of multicriteria decision making in drug discovery is addressed.

# 2 Overview of Drug Design

## *2.1 Process Overview*

The entire process of drug discovery and development can be depicted as a rocketship with stages, an image that portrays the "funneling" effect as fewer compounds are under consideration at successive stages of the process. The focus of this chapter is screening issues in lead generation (the first stage of the rocket) which begins with the chemical entity and the biological target. The chemical entity (compound) may be a small molecule, a peptide, or a large protein. Typically, chemical entities are purchased from external providers or synthesized within a company. The compound can be viewed as binding or docking to the biological receptor or target in order to competitively inhibit, or else to induce, some biological signal such as the production of a protein or hormone, resulting in a specific response. The number of molecules tested is dependent on reagent costs and other practical factors. This chapter adopts the following paradigm for drug discovery and development.

1. A target is validated to establish a direct link (such as a gene or a process in the body, or a virus or a parasite) to the disease state of interest and the feasibility of controlling the target to obtain the desired therapeutic benefit. This stage involves scientific study that can be catalyzed by genomic and proteomic technologies. Target validation requires careful scientific experiments designed to explore how a target influences a biological response or disease state.
2. A high-throughput screen (HTS) is designed, optimized, calibrated, validated, and run to obtain biological response data at a single concentration for

200,000 compounds (Section 4). This may involve whole cells, enzymes, or other in vitro targets. Reducing variability is crucial. Sittampalam et al. (1997) introduced the concept of *signal window*, a method whereby two controls are used to diagnose the ability of the assay to distinguish between actives and inactives in the presence of background noise.

3. The most promising compounds, or *actives*, typically numbering from 1000 to 5000, are then tested in a *secondary* screen which involves testing each compound at 5 to 10 different concentrations. These results are modeled with a nonlinear dose–response curve and for each molecule a summary measure is computed such as a 50% inhibitory concentration (IC50) or a 50% efficacious concentration (EC50).
4. The secondary assay reduces the set of actives to those for which potency reaches at least 50% of the maximum potency of a reference compound, at some concentration. Typically there are 500 to 1000 of these compounds, and they are called *hits*. Many hits may be nonspecific or for other reasons may offer no prospect for future development. (In subsequent sections the distinction between *active* and *hit* is blurred.)
5. The hits are examined in a series of careful studies in an effort often called *hit to lead*. Chemists look at the hits and classify them into four to eight broad series and, within each series, they try to find a structure–activity relationship. The chemists characterize these SARs by testing (in the secondary assay) a few hundred or a few thousand more molecules, thus expanding each SAR. Out of these SARs, the chemists and biologists choose a few hundred compounds to be tested in cell-based or enzyme in vitro screens. These screens require careful design and validation. From the molecules run through the in vitro testing, 100 or so may go through in vivo single-dose tests using a rodent or some other animal model. Some 10 to 40 of these molecules are finally tested for in vivo efficacy in a full dose–response experiment performed on the animal of choice.
6. The lead compounds undergo careful studies in an effort known as *lead optimization*. At this point any remaining issues with metabolism, absorption, toxicity, and so on, are addressed through molecular modification.

Some research groups contend that the HTS step should be eliminated and replaced with a number of rounds of iterative medium-throughput screening (Section 7). It is an issue of quantity versus quality. The lower-throughput screens tend to have lower variability ("noise") and less dependence on the single concentration test as an initial triage. The iterative approach is closely akin to a strategy in machine learning (Section 6) known as *Active Learning*.

In a successful project, the steps outlined above will lead to a *First Human Dose* (FHD) clinical trial. How well those prior steps are done will, in part, determine the success or failure of the human clinical trials. The adoption of high-throughput screening and combinatorial chemistry methods in the early 1990s offered promise that a shotgun approach to drug discovery would be possible. It was soon learned that simply increasing the volume of screening results cannot be the answer. The number of potential chemical entities is staggering, being estimated to be between

$10^{20}$ and $10^{60}$. The efficient exploration of a landscape of this magnitude requires prudent use of machinery, human expertise, informatics, and, even then, an element of fortuity. On average, for every 5000 compounds that enter a hit-to-lead phase, only five will continue on to clinical trials in humans, and only one will be approved for marketing. It is analogous to searching for a small needle in a whole field of haystacks. In these terms, the future of drug design lies in no longer searching for the needle but, instead, constructing the needle using available clues.

## 3 False Negatives and False Positives

In primary screening, compounds are tested at a single concentration; those whose response exceeds a prespecified threshold are labeled as "active" and the rest as "inactive". Typically, 200,000 compounds are screened, giving numeric potency results for each, then, based on exceeding a threshold, about 2000 are labeled as active and 198,000 as inactive. The actives are studied further at multiple concentrations and the inactives are henceforth ignored. A *false positive* error occurs when a compound labeled as active is, in fact, inactive when studied in the more careful multiple concentration assay. The false positive rate can be lowered by raising the decision threshold, or "hit limit", but at the cost of increasing the false negative error rate. In most HTS screens, of those compounds flagged as active in a primary screen, roughly 30% to 50% are found to be inactive in the multiple concentration–response follow-up study.

A *false negative* error occurs when a compound that is actually active is not labeled as active. Biological noise, for example, and the choice of hit threshold can affect the false negative error, as well as mechanical failures such as a leaking well. Mechanical failure errors are unrelated to the true potency of the molecule. The false negative rate is unknowable because the vast majority of compounds are not studied further, but it can be estimated from small studies. From past HTS screens at Lilly, we have estimated that a mechanical failure false negative occurs in roughly 7% to 12% of compounds in an HTS screen, with a total false negative error rate ranging from 20% to 30%. Aside from the mechanistic type of false negative, the false negative rate can be viewed as a function of the activity level—the greater the activity of the molecule, the lower the chance of a false negative error.

Experimental results from an HTS assay are not the "truth" but merely an estimate of the true potency of a molecule. Because molecules are only measured one time in the HTS setting, there is a high degree of uncertainty. One thing that can be done is to look for highly similar molecules and treat them as pseudo replicates of the same "parent" molecule. See Goldberg (1978) for further discussion on estimating the error rate.

One effective way of dealing with experimental errors is to build a predictive model and score the screening results through the model, then to look at discrepancies between the experimental screening result and the model prediction. Often the highly potent but mechanical failure type of false negatives or false positives

TABLE 1. Breakdown of hit rates from the screening follow-up results.

| Compound source | Hit rate |
|---|---|
| 150,000 original compounds | 3% |
| 2050 new compounds | 34% |
| 250 potential false negatives | 55% |

can be identified. In practice, the false positives will be tested in the secondary screen and found to be inactive when that screening result is observed. Some researchers favor raising the threshold to reduce the number of compounds labeled as active. The false negatives can then be identified by a statistical or predictive model and rescreened. In one recent project at Lilly we followed up a 150,000 compounds HTS with a small library (that is, a small collection of molecules) of 2300 compounds. A predictive model was trained (or fitted) using the 150,000 primary results and used to select 2050 molecules that had not yet been tested. The same model was used to identify 250 molecules that were screened in the 150,000 and found to be inactive, yet scored by the model as highly active. These 250 were the potential false negatives that were to be retested. Fully 55% of these false negatives were active upon retesting. The breakdown of hit rates is given in Table 1.

The 3% hit rate in the primary screen was a concern, as such a high number suggests a problem in the assay. It was found that there was a carryover problem in which sticky compounds were not being completely washed from the robotic tips. Such a trend can be found easily by analysis of hit rate as a function of well location. This problem was resolved before the secondary runs (rows 2 to 3 of the table) were made.

A computer experiment was done to confirm and further explore the above findings. An iterative medium-throughput screening operation was simulated with different levels of false negative rates, reflecting historical error rates seen across past screens. For each level of false negative rate, the appropriate proportion of true actives was randomly chosen (from a nonuniform distribution that is a function of the "true" activity level of the molecule, based on historical data) and relabeled (incorrectly) as inactive. The model was trained on this "polluted" data set and used to select the set of compounds for the next round of testing. Computer experiments of this type can be run many times to explore the behavior of the predictive models under realistically stressed circumstances. For this particular experiment, the model was able to find 25 times more false negatives than a random (hypergeometric) search would produce, up to a false negative rate of 30% at which point the enrichment over random decreases from 25-fold to about 15-fold higher than random.

In both screening and predictive modeling, the relative cost of false negatives versus false positives is an important component of decision making. From a business standpoint, false negatives represent lost opportunities and false positives represent wasted efforts chasing down "red herrings." The resources wasted

chasing false positives is particularly troubling. The current trend is to reduce the false positives and to tolerate the increased number of false negatives that results.

## 4 Molecular Diversity Analysis in Drug Discovery

In the last two decades, three technologies have been co-developed that enable a significant shift in the process of lead generation:

$$\begin{aligned} \text{Combinatorial Chemistry} &\Rightarrow \text{Large libraries of molecules} \\ \text{High-Throughput Screening} &\Rightarrow \text{Many biological data points} \\ \text{Cheminformatics} &\Rightarrow \text{Many molecular descriptors} \end{aligned}$$

The adoption of high-throughput screening and combinatorial chemistry methods in the early 1990s led to an immense interest in molecular diversity. It was widely expected that making diverse libraries would provide an increase in the number of hits in biological assays. It took a while to realize that this was the wrong expectation. Molecular diversity designs do offer great benefits, but more in the enhancement of the quality, rather than quantity, of information from a screen. It became clear that other properties of molecules, beyond mere structural novelty, need to be considered in screening. This led to extensive work on "drug-likeness" and an attempt to achieve a balance between diversity and medicinal reasonableness of molecules.

### *4.1 Molecular Diversity in Screening*

Molecular diversity analysis is useful in several contexts:

- Compound acquisition: this avoids purchasing a compound very similar to one that is already owned.
- General screening for lead identification: screening a diverse library is a sensible approach when little or nothing is known about the target or possible lead compounds.
- Driving an SAR effort away from prior patent claims.

The second context, general screening, involves selecting subsets of molecules for lead generation. Experimental designs are considered because it is not feasible to screen all molecules available. Even with the application of high-throughput screening, the demand for screening outpaces the capacity. This is due to the growth of in-house chemical databases, the number of molecules synthesized using combinatorial chemistry, and the increasing number of biological targets fueling discovery projects. In addition, novel screens that are not amenable to HTS automation may be attractive from a competitive standpoint but the gap between screening capacity and screening opportunities in this case is particularly daunting. Given this imbalance, methods for selecting finite subsets of molecules from potentially large chemical databases must be considered. Possible selection

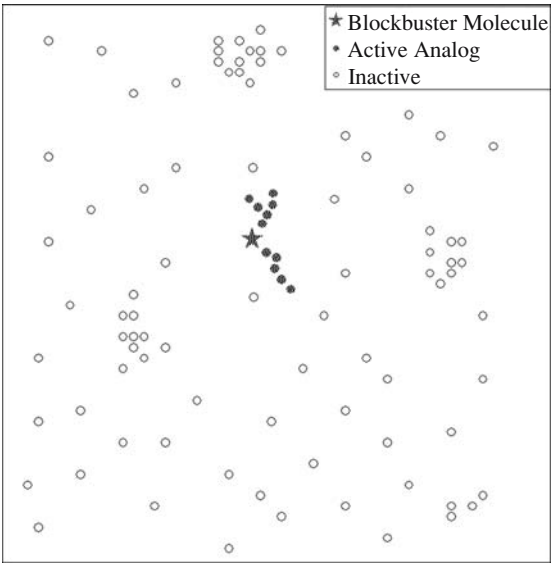



FIGURE 1. SAR paradigm. Fictitious two-dimensional projection of the property space of feasible druglike molecules.

strategies include: random, diverse, and representative selection, each of which may be performed as a biased or directed analysis if information such as a drug-likeness score is available to weight the analysis in favor of certain classes of molecules. A requirement of every selection method considered here is its computational feasibility for the databases of hundreds of thousands to millions of compounds that are now common with the application of combinatorial synthesis. For example, many distance-based selection strategies involve computation and storage of all pairwise distances for molecules in a database. If the number of molecules $n$ is 300,000, then there are approximately $45 \times 10^9$ (calculated from $\binom{n}{2}$) distances to compute and/or to store. This is a formidable task, necessitating creative computational solutions.

Figure 1 illustrates a modern paradigm of drug hunting processes. In this fictitious two-dimensional projection of the space of feasible druglike molecules, open circles represent compounds that are not active relative to a specific target, solid circles are active compounds, shown in one contiguous series of related molecules, and the star is the blockbuster drug that is still undiscovered. In a primary screen, finding a single solid circle is all that is needed. The medicinal chemistry teams can follow up by making systematic changes to any one of the active compounds to explore the whole series and find (or invent) the blockbuster drug. An important point is that the blockbuster may not exist in the corporate collection. A typical lead generation or lead optimization project involves not only testing molecules in current libraries, but also synthesis of new molecules. Molecular modification and subsequent testing is the way the trail gets blazed, through characterizing the SAR.

In recent years there has been strong dogma contending that, in filling a fixed screening capacity, it is important to screen "backups", that is, molecules that are closely related. This argument is motivated by the high rate of false negatives in primary screening. Thus screening two or more compounds from the same related series effectively gives pseudo replicates. If one compound turns out to be a false negative, it is likely that another from the same series will screen as positive and thus, the active series will not be missed. This rationale is popular in the industry. However, at Lilly we have demonstrated, through both simulations and retrospective analysis, that it is better to tolerate the false negatives in favor of sampling a larger number of different series. The motivating principle for this position is that testing two closely related compounds (or *analogues*) is often equivalent to testing the same hypothesis twice, which comes at the expense of testing a different hypothesis; see Wikel and Higgs (1997).

Optimizing molecular diversity has the potential to maximize the information gained about an SAR in the early stages of screening. Suppose a random screening gives the same number of hits as a diverse screening. Then one would favor the diverse set of hits, because this increases the chance of at least one structural lead with a favorable ADME and toxicity profile. In fact, for primary screening, it is often better to have 10 novel hits than 200 hits that are analogues of each other. The proper focus, in our view, is quality of information gleaned from the screen.

Most pharmaceutical companies have clusters of compounds (for example, Lilly has many SSRIs, cephalosporins, and so on). There are many analogues clustered tightly in local subregions of chemical space, reflecting historical SARs investigated around related targets. A random sample will reflect the densities of the compound classes in the collection; thus testing a random sample of molecules for a certain biological activity will be equivalent to testing the same hypothesis many times over.

## 4.2 Descriptors

Computationally, a *structure space* (represented as a set of two-dimensional graphs of molecule structures) is mapped to *property (or chemical) space* ($\mathcal{R}^p$) (for example, Figure 2), where each point is a vector of values of each of $p$ variables,

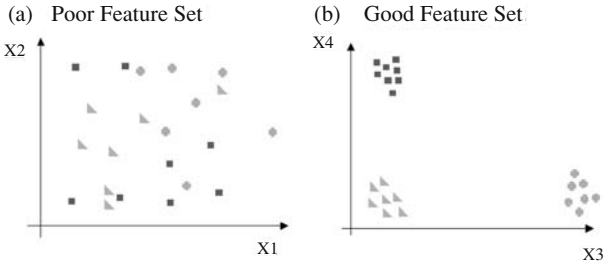



FIGURE 2. Descriptor validation example: (a) poor feature set; (b) good feature set.

called *descriptors*, or sometimes *properties*. The descriptors can be binary, integer counts, or continuous variables. A descriptor may be as simple as a count of the number of oxygen atoms in a molecule, or as sophisticated as an estimate of the three-dimensional polar surface area around the molecule. Molecules are assigned positions in this high-dimensional descriptor space through their properties. The relationships defining their molecular diversity are, therefore, represented through their coordinates or positions in this space. The distance metrics most often used are Euclidean and Mahalanobis for properties, and Tanimoto (Jaccard) for binary bit strings; see Section 4.5.

Prior to selecting a set of molecules from a database, it is often necessary to preprocess the molecular descriptors to replace missing descriptor values and to scale the descriptors. Although it is possible to develop distance metrics that are tolerant to missing values, at Lilly we have focused on imputing (replacing) missing values and using distance metrics that assume all descriptor values are present.

A set of molecules is commonly described with anywhere from 4 to 10,000 descriptors. It is also possible to represent molecules with sparse descriptors numbering up to 2 million. Variable selection, or descriptor subset selection, or descriptor validation, is important, whether the context is supervised or unsupervised learning (Section 6).

## *4.3 Molecule Selection*

Discussions about molecular diversity involve the concepts of "similarity" and "dissimilarity" and may be confusing as their meanings are content related. Similarity is in the eye of the beholder. Chemists may find similarity hard to define, but they generally are quick to identify it when they see it and at times are willing to debate the similarity of one structure to another. Similarity is not absolute, but relative to the space in which it is defined. In chemistry, this means the definitions must always be held in context to the property space used to define the structures.

If characteristics are known about the biological target then this target-specific information may be used to select a subset of molecules for biological testing. Various database searching methods, molecular similarity methods, and molecular modeling methods could be used to identify a favored (or biased) subset of molecules for biological testing. This corresponds to the second row of Table 2. One example of this situation is the case of neuroscience targets. If very little

TABLE 2. Subset selection strategies for primary screening at Lilly.

| Situation | Strategy |
|---|---|
| No target information | Diversity Selection |
| Expert judgment or literature information about target | Directed Diversity Selection |
| Experimental results related to target, or structure of target known | QSAR, Predictive Modeling |

is known except that the receptor of interest is in the brain, a biased diversity selection would be more useful than an unbiased one. For example, one might construct a weight function based on the number of positively charged nitrogen atoms in a molecule, because this is often observed to be present in desirable neuroscience drugs. If there are no positively charged nitrogen atoms, or if there are more than two, the weight function is very low and otherwise very high. Other factors related to toxicity, solubility, and other aspects of medicinal viability of molecules could be included in the weight function. Then a weighted diversity selection could be performed to construct a reasonable starting set for initial screening.

## *4.4 Descriptor Validation and Variable Selection*

The concept of molecular similarity is strongly linked with the "SAR Hypothesis" of Alexander Crum-Brown (Crum-Brown and Fraser, 1869) which states that compounds that are similar in their structure will, on average, tend to display similar biological activity. A modest extension holds that one can build mathematical models from the numerical descriptors to describe a relationship between the chemical structure and the biological activity. When chemists discuss similarity of two molecules, they often make arguments about the biological effects or binding potential of the compounds. There is a concept of *bioisostere* which says that some chemical fragments function in the same way as other chemical fragments (for example, a sulfur may behave like a methyl group). An ideal set of molecular descriptors would be one that contains properties characterizing all aspects important to potency, selectivity, ADME, and toxicity. Because our understanding of any one of these processes is limited, expert judgment is needed. Descriptors considered generally important include those describing lipophilicity, molecular shape, surface area and size, electronic properties, and pharmacophoric elements such as hydrogen bond donors and acceptors.

Just as with variable subset selection in linear regression, there are risks akin to over-fitting a training set. (A *training set* is the subset of data used to fit the model.) The topic of how best to do descriptor validation has been hotly debated, and numerous ideas have been proposed, but the general goal is to select that subset of descriptors that best achieves some sense of separation of the classes of compounds, as illustrated in Figure 2. This figure illustrates three structural series in two hypothetical two-dimensional configurations. Descriptors $x_3$ and $x_4$ are more useful because they separate the different structural classes.

There are dimensionality issues. Later we propose Mahalanobis distance (Section 4.5) as a good metric for diversity analysis. With $p$ descriptors in the data set, this metric effectively, if not explicitly, computes a covariance matrix with $\binom{p}{2}$ parameters. In order to obtain accurate estimates of the elements of the covariance matrix, one rule of thumb is that at least five observations per parameter should be made. This suggests that a data set with $n$ observations can only investigate approximately $\sqrt{2n/5}$ descriptors for the Mahalanobis distance computation. Thus, some method for subset selection of descriptors is needed.

In the case of molecular diversity, there is no response to guide the variable subset selection (unsupervised learning) and hence creative ways to do subset selection are needed. Ideally, chemical similarity is defined by the target or receptor. If one has information about the target/receptor then it is more useful to do QSAR (last row of Table 2). If such information is lacking, then one must impose an arbitrary definition on chemical similarity in order to avoid testing duplicate, or very similar, hypotheses in screening. Thus, at Lilly, our molecular diversity tools are generic and depend on a generic notion of similarity that is relatively independent of biology (rows 2 and 3 of Table 2).

## *4.5 Distance Metrics*

Distance-based methods require a definition of molecular similarity (or distance) in order to be able to select subsets of molecules that are maximally diverse with respect to each other or to select a subset that is representative of a larger chemical database. Ideally, to select a diverse subset of size $k$, all possible subsets of size $k$ would be examined and a diversity measure of a subset (for example, average near neighbor similarity) could be used to select the most diverse subset. Unfortunately, this approach suffers from a combinatoric explosion in the number of subsets that must be examined and more computationally feasible approximations must be considered, a few of which are presented below.

Given two molecules $a$ and $b$, let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote their vectors of descriptors. The Mahalanobis distance between $a$ and $b$ is defined as:

$$d(a, b) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{V}^{-1} (\boldsymbol{x} - \boldsymbol{y})},$$

where $\boldsymbol{V}^{-1}$ denotes the inverse of the covariance matrix, $\boldsymbol{V}$, of the vectors of the descriptor values of all the molecules. If $\boldsymbol{V} = \boldsymbol{I}$ the result is Euclidean distance:

$$d(a, b) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2},$$

where $x_i$ and $y_i$ are the $i$th elements of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

The effect of the $\boldsymbol{V}^{-1}$ is to divide each descriptor by its standard deviation, so that some descriptors do not dominate others due to mere differences of scale. Many cheminformaticians compute the standard deviations explicitly, but this alone is not sufficient. The off-diagonal elements of the inverse covariance matrix adjust for overweighting (due to high correlations between descriptors) of latent aspects of a molecule, such as size.

A common practice is to scale each descriptor to have standard deviation of 1. Another is to compute principal components and confine the analysis to the first $h$ components, where $h$ may range from 1 to 20. This is an ad hoc form of dimension reduction that does not remove irrelevant information from the analysis. At Lilly, we prefer a careful descriptor validation to avoid including many irrelevant descriptors into the analysis, combined with a dimension reduction criterion using

the $\sqrt{2n/5}$ rule of thumb, followed by a Mahalanobis distance computation using all the descriptors that remain.

For presence or absence of features in the molecules, represented by binary bit strings $\boldsymbol{x}$ and $\boldsymbol{y}$ as descriptors, the Tanimoto coefficient is a popular metric for similarity:

$$sim(a, b) = \frac{(\text{bits on in both } \boldsymbol{x} \text{ and } \boldsymbol{y})}{(\text{bits on in } \boldsymbol{x}) + (\text{bits on in } \boldsymbol{y}) - (\text{bits on in both } \boldsymbol{x} \text{ and } \boldsymbol{y})}.$$

Then

$$d(a, b) = 1 - sim(a, b).$$

Now consider $d(a, b)$ to be a generic distance metric of which Tanimoto, Euclidean, and Mahalanobis are three cases. Then, the distance between molecule $a$ and the set of molecules $B$ is defined as follows,

$$d(a, B) = \min_{b \in B} d(a, b),$$

and the overall dissimilarity of a set of molecules $M$ is defined as

$$dis(M) = \frac{1}{n} \sum_{a \in M} d(a, M \backslash a), \tag{1}$$

where $M \backslash a$ denotes the set $M$ with the molecule $a$ removed.

These metrics are used by design algorithms for selecting dissimilar molecules for chemical analysis (see Section 5.2).

## 5 Subset Selection Strategies

A requirement for any subset selection method is the ability to accommodate a set of previously selected molecules, where augmentation of the pre-existing set is desired. For example, when purchasing compounds, the goal is to augment what is already owned so that the current corporate collection would be used in the analysis as the pre-existing set of molecules. The goal then is to select a subset of the candidate molecules that optimizes a specified criterion with reference to the molecules in both the candidate set and the previously selected set.

In the case of iterative medium-throughput screening, at any given point in the process, the set of molecules that have been screened thus far is the previously selected set for the next round of screening. In choosing molecules for the next iteration, one may have a selection criterion such as predictive model scores but a diversity criterion may also be applied: it is not desirable to screen something identical, or nearly identical, to that which was screened in previous rounds.

There are two main strategies developed to select diverse and representative subsets of molecules, namely, cell-based methods and distance-based methods.

## 5.1 Cell-Based Methods

Cell-based methods divide the space defined by a set of molecular descriptors into a finite set of "bins" or "buckets". Each molecule is then assigned to one of the bins. Structurally similar molecules will occupy the same or adjacent bins and dissimilar molecules will occupy bins that are spatially well separated. A diverse subset of molecules can be identified by selecting a single molecule from each of the occupied bins. Databases can be compared by examining the occupancy of bins with molecules from different sources. For example, commercial databases such as Comprehensive Medicinal Chemistry (2003), World Drug Index (2002), and Maccs Drug Data Report (2003) contain molecules that can be used to define the historically medicinally active volume (bins) of chemical space. Compounds in another database, or collection, that fall within the bins defined by these databases can then be selected for biological testing.

Cell-based methods have the advantage that they are intuitive and computationally more efficient than many distance-based methods. However, cell-based methods suffer from a problem known as the "curse of dimensionality." Consider a database with each molecule described by 20 molecular descriptors. Subdividing each molecular descriptor into merely 5 segments (or bins) will result in $5^{20}$, or approximately $10^{14}$ bins. Even with large chemical databases, most of the bins will be empty and many bins will contain a single molecule. Outliers wreak havoc. Just one molecule whose molecular descriptors take on extreme values will cause the majority of molecules to be allocated to a small number of bins. In either case, a cell-based method will present problems in selecting a diverse subset of molecules. Thus, cell-based methods require a significant reduction in dimensionality from the many possible molecular descriptors, attention to outliers, and careful consideration of how to subdivide each dimension. An application to drug discovery screening, which addressed the issues of outliers and dimensionality, was applied to large databases by Cummins et al. (1996).

## 5.2 Distance-Based Methods

Statistics has a long-standing role in design of experiments. There is a long history of the use of information optimal designs (for example, D-optimal designs), which consist of the most informative points and are useful in designed experiments where the "true" model is known. *Space filling* designs are used in numerous contexts including geographical modeling (literal space filling), modeling response surfaces, multivariate interpolation, and chemical library design.

A more in-depth discussion of three selection methods that are computationally feasible with very large chemical databases is now given to highlight the issues that must be considered when applying many of these molecular diversity selection

methods. The three design methods described here are edge, spread, and coverage designs. Each design method optimizes a specific objective.

- Edge design objective: obtain minimum variance estimates of parameters in a linear model.
- Coverage design objective: select a subset of molecules that is most representative of the entire library. Heuristically, the distance from the chosen subset to the remaining candidate points should be small. One might imagine a set of umbrellas positioned to cover as many candidate points as possible.
- Spread design objective: select the maximally dissimilar subset of molecules. This requires maximizing the distance of points within the subset from each other. One analogy for this is electron repulsion.

Edge designs are often constructed using D-optimal design algorithms. Molecules selected using D-optimal designs populate the edge of descriptor space by first filling in the corners and then moving around the boundary. Edge designs are appropriate when one intends to fit a linear regression model where the descriptors are the predictors in the model, for example, if biological activity is modeled as a function of the descriptors. This is usually the situation in lead optimization, rather than lead generation.

Spread and coverage designs are space-filling designs. Let $C$ be the *candidate set*, that is, the set of possible design points. Once the criterion (*space filling*) is well defined, selecting the points $M \subset C$ to be space filling is simply an optimization problem.

The objective of a spread design is to identify a subset of molecules in which the molecules are as dissimilar as possible under a given similarity metric. For a given metric to measure the similarity of a subset, all subsets of size $k$ (plus any molecules previously selected) could be evaluated and the subset that produces the lowest similarity measure chosen. In practice, simple non-optimal sequential algorithms are often used to approximate the maximally dissimilar subset: two such algorithms are described below.

1. Maximum Spread Algorithm

The goal: out of all $\binom{n}{k}$ subsets of $k$ molecules from a candidate set $C$, find the subset $M^*$ where $dis(M^*)$, defined in (1), is largest. The problem is that it is not feasible to enumerate and evaluate all possible subsets. The solution is to use a sequential approximation (greedy algorithm).

a. Select the first compound from the edge of the design space.
b. Select the second compound to be most distant from the first.
c. Select subsequent compounds in order to maximize the minimum distance to all previously selected compounds.

This is the algorithm proposed by Kennard and Stone (1969). At Lilly we have focused on an efficient implementation of this approach applied to large chemical databases and have not implemented design optimization due to the marginal

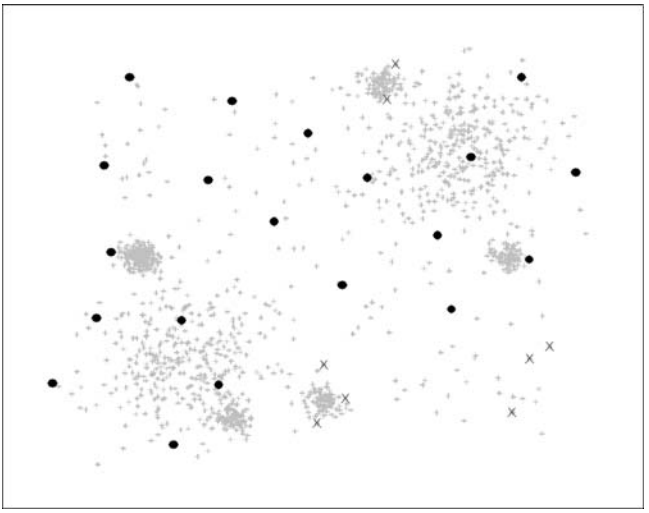


FIGURE 3. Spread design for fictitious example. (Reproduced from Higgs et al., 1997 with permission.)

design improvements and increased computational time. To illustrate, the SAS (2003) OPTEX (*CRITERION* = *S*) procedure was used to select 20 points from the 1400 two-dimensional points shown in Figure 3 using a modified Fedorov optimization algorithm (Cook and Nachtsheim (1982)). The OPTEX procedure seeks to maximize the harmonic mean distance from each design point to all other design points. Eighty different designs were generated using the sequential method of Kennard and Stone and compared with those obtained by the modified Fedorov optimization method. On average, the Fedorov optimization generated a design that was 8.5% better than that obtained from the simple sequential selection method but required eight times more computational time. In larger data sets of 200,000 or more compounds this can mean a choice of eight hours versus three days to find a design.

2. Maximum Coverage Algorithm

Define the coverage of a set $M$, where $M \subset C$ as:

$$\text{cov}(M) = \frac{1}{n} \sum_{\alpha \in M} d(a, C \backslash M),$$

where $C \backslash M$ is the set $C$ with the set $M$ removed. The goal: out of all $\binom{n}{k}$ subsets of $k$ molecules with descriptor vectors in $C$, find the subset $M^*$ where $\text{cov}(M^*)$ is smallest. This is often approximated using cluster analysis (see Zemroch, 1986).

In Section 5.3 the different design types are compared.

## 5.3 Graphical Comparison of Design Types

Figures 3–5 show a fictitious two-dimensional data set reproduced from Higgs et al. (1997) with permission. The data set contains 1400 hypothetical molecules and is constructed to illustrate the differences between edge, spread, and coverage designs. The data set was constructed to have five tightly packed clusters (bivariate normal), two loosely packed clusters (bivariate normal), and molecules uniformly distributed over the two-dimensional design space. For illustrative purposes, eight molecules were randomly chosen and labeled with an "X" as having been selected in a previous design. Future selections should complement these eight molecules. The data were simulated in two dimensions to depict how a pharmaceutical compound collection might appear in some two-dimensional projection. Certain regions are sparse with low density whereas other regions are highly clustered, reflecting the synthetic legacy of the company.

Figure 4 shows 20 molecules selected using an edge (D-optimal) design to augment the previously selected molecules. Two quadratic terms and one linear interaction term were included in the model used to select this design in order to force some interior points into the selection. Figure 5 shows 20 molecules selected using a $k$-means clustering approximation to a coverage design to augment the previously selected molecules. Figure 3 shows 20 molecules selected using the Kennard and Stone approximation to a spread design (see, for example, Johnson et al., 1990) to augment the previously selected molecules.

Although not shown in the figures, a random selection is often considered the baseline method of subset selection. Random sampling typically selects many molecules from the dense clusters, and several molecules near the previously selected molecules. Spread designs select the most diverse subset of molecules

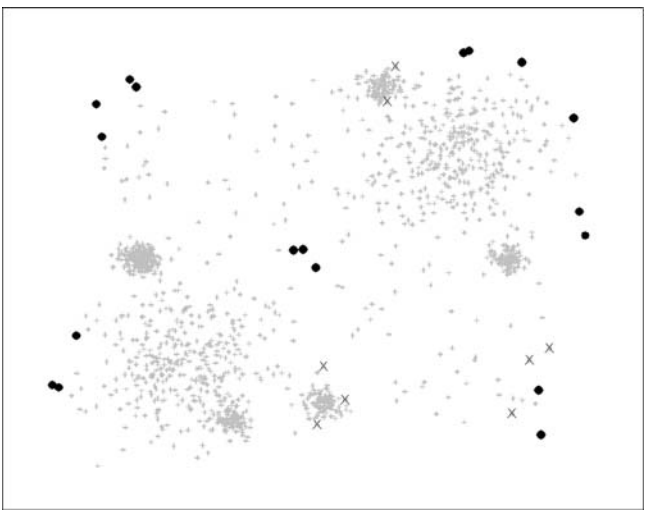


FIGURE 4. Edge design (D-optimal with interactions) for fictitious example. (Reproduced from Higgs et al., 1997 with permission.)

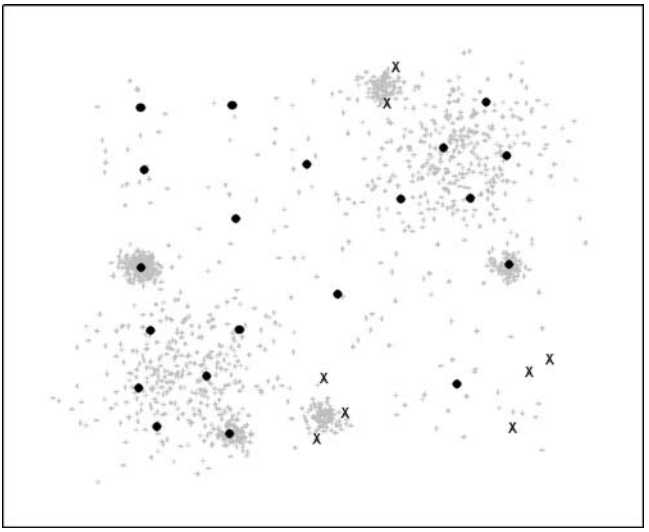


FIGURE 5. Coverage design for fictitious example. (Reproduced from Higgs et al., 1997 with permission.)

(relative to the other methods presented here), including molecules near the edges as well as throughout the design space. Spread designs ignore the density of the candidate points and focus rather on efficient exploration of the *space* populated. Coverage designs select molecules near the center of clusters. Molecules near the edges of the design space are naturally avoided because they are unlikely to be near the center of a cluster.

## 5.4 Combinatorial Chemistry Example

This example illustrates the usefulness of a tool that assigns a rank ordering to molecules in a set. A combinatorial chemistry collection at Lilly consisted of a number of separate libraries. The question arose as to which of the libraries was the most diverse. To answer this question, a spread design was used to rank the combinatorial molecules. We pooled 22 combinatorial libraries (105,640 molecules) with a set of 32,262 corporate library molecules. We rank ordered the combinatorial molecules relative to the corporate library molecules; that is, the corporate library molecules were marked as pre-selected and the task was for the combinatorial candidates to augment them as well as possible. The spread design chose molecules from the pool irrespective of which library they came from—the only criterion was their diversity. We examined the cumulative number of molecules selected from each combinatorial library as a function of spread design rank, as follows. The first molecule chosen was the one most dissimilar to the corporate collection and received a rank of 1. The next molecule was that which was most dissimilar to both the corporate collection and the first molecule, and received a rank of 2, and so on. Libraries that were drawn from most frequently by the algorithm in

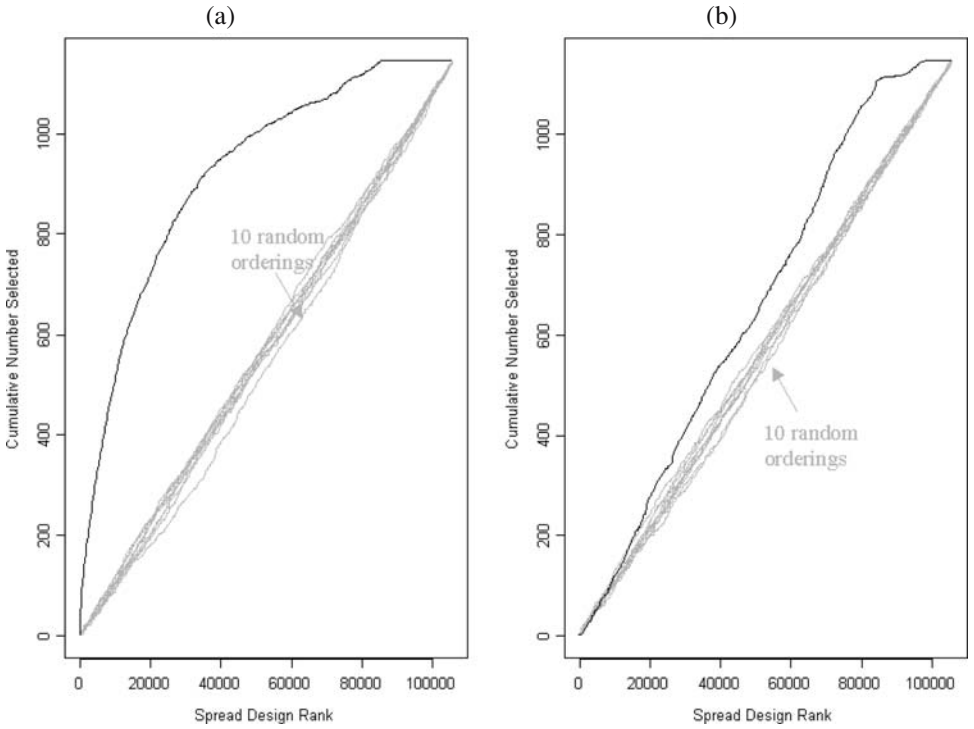



FIGURE 6. Combinatorial libraries comparison: cumulative number of molecules selected versus the rank of the spread design for each of 10 random orderings of molecules within (a) Library A; (b) Library B.

the early stages (early ranks) were taken to be the most diverse libraries. In the case of Figure 6, library A was far better than library B at augmenting the current collection.

This example shows how spread designs can be used to solve practical problems. There is always a descriptor selection problem, as chemists continue to invent new molecular descriptors. Which should be used? Which molecular similarity measure performs best? Controlled experiments are expensive. Simulation can be used as a guide.

All of this effort is invested in the first of a number of iterations in the drug discovery cycle and the later stages are much more rewarding. At Lilly, we move as quickly as possible into the predictive modeling stages.

## 6 Machine Learning for Predictive Modeling

*Machine learning* is defined as the use of algorithms to generate a model from data, which is one step in the knowledge discovery process, applied in the context

of QSAR (last row of Table 2). The last decade of machine learning advances has seen tremendous increases in prediction accuracy, largely due to model averaging techniques. A good starting point for reading about such ensemble methods is the paper of Breiman (1996) and a valuable discussion about algorithmic versus parametric modeling approaches is provided by Breiman (2001b). Hastie et al. (2001) gave a broad overview of statistical learning (see, especially, the figures on pages 194 and 199). Predictive models can serve as useful tools and have made substantive contributions to many disciplines.

## *6.1 Overview of Data Handling and Model Building Steps*

Figure 7 gives a brief layout of sequential steps for a typical data modeling exercise. The first step, which is by far the most time consuming, starts from a representation of the structures of the molecules and ends with a "training set" of descriptors to be used in the model selection step. Medchem filtering, in step 1, is an application of expert judgment to chemical structural data. Certain fragments of molecules are known to be highly reactive, or carcinogenic, or unstable, or otherwise undesirable, and these molecules can be eliminated at this first step with a simple rule-based algorithm. Data cleaning is, by far, where most of the time is spent.

The data cleaning steps may involve the removal from the data frame of columns (of descriptor values) that are constant or nearly constant, imputing missing values and eliminating columns that are redundant due to a strong relationship with other columns. All these steps are easily automated. Approximate algorithms can easily be developed that are more than 100-fold faster than those available in commercial packages.

The next step of data cleaning is to perform a replicate and pseudo replicate analysis of the experimental values. When replicate data are available, highly discrepant results can point to problems with the experimental data. When replicate results are not available, pseudo replicates are almost always present in the data. Often the same chemical structure exists more than once in the results file, where the different identifiers refer to different batches of the same material. Thus, a

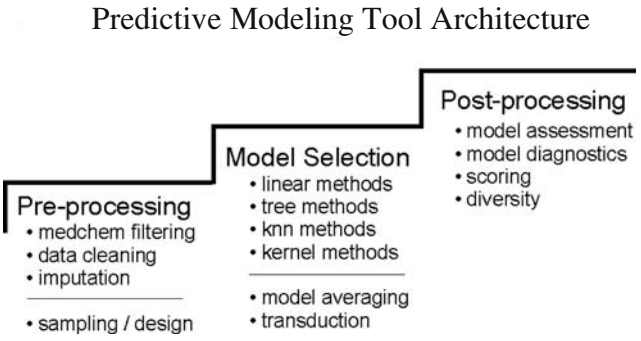



FIGURE 7. Steps involved in predictive modeling.

large discrepancy in the biological response for two such identifiers suggests that a follow-up meeting with the appropriate biologist is needed in order to resolve the experimental discrepancies.

Another aspect of data cleaning arises when data come from different laboratories. Then one is faced with the task of placing the results in a reliable and consistent context (a sort of "metaanalysis"). Another data cleaning task involves the imputation (estimation) of missing values. Often the programs that compute descriptors will fail on unusual molecules and then those molecules are usually removed from further consideration. However, sometimes a failure is not a reflection of the desirability of the molecule and imputation of the missing values is then a reasonable strategy.

The final portion (the sampling step) of the first step of Figure 7 is to create a *training set* which is the set of data to be used for fitting the model in the model selection stage. This may be done in several different ways. The simplest is merely to take random subsamples of the data and to split the data into training and test data. A more rigorous approach involves splitting the data into a training, validation, and test set (Hastie et al., 2001, pages 195–196), where the test set is used only once for assessing the error of the final model (after the model selection studies have finished) and the training and validation sets are used for model selection to compare competing modeling methods.

The "design" question of what proportion of the data to use for training, relative to the test set, is an important one. If the test set is too small, the estimates of error will be unreliable, highly variable, and likely to have a high downward bias. On the other hand, if the training set is too small the estimates of error will have high upward bias.

The second step shown in Figure 7 lists the actual model training steps. These typically involve a model selection exercise in which competing modeling methods are compared and a choice of one or more modeling methods is made. Listed in the figure are four of the many popular classes of modeling approaches. We use all of the methods listed; see Section 6.3.

## *6.2 Error Rates*

Some of the examples and discussion in this chapter draw on the two-class classification problem, which here is "hit" versus "inactive". The word "active" refers to a validated hit, that is, a molecule that truly does exhibit some level of the desired biological response. A key point is that an assay is itself an estimator. With this in mind, definitions and a discussion of error rates are given in the context of predictive models. Borrowing from the terminology of signal detection, the "sensitivity" of a model refers to the fraction of observed hits that are classified as (or predicted to be) hits by the model, and "specificity" refers to the fraction of observed inactives classified as inactives by the model. An observed hit is not necessarily an active molecule, but simply a molecule for which the primary screening result exceeded a decision threshold. Whether such a molecule turns out to be an active is a problem that involves the sensitivity of the assay, but the task at hand is for

a model to predict accurately the primary screening outcome and to assess the accuracy of the model for that purpose.

Let $\hat{I}$ denote "predicted by the model to be inactive" and $I$ denote "observed to be inactive in the assay by exceeding the decision threshold", with analogous definitions for $\hat{A}$, "predicted to be a hit", and $A$, "observed to be a hit". With the null hypothesis that a compound is inactive, we have:

$$\begin{aligned} \text{specificity} &= P(\hat{I} \mid I) = P(\text{model prediction} - \mid \text{observed} -) \\ P(\text{Type I error}) &= P(\hat{A} \mid I) = P(\text{false positive}) = 1 - \text{specificity}. \end{aligned}$$

Similarly, 1 minus the sensitivity gives the probability of Type II error or the false negative rate:

$$\begin{aligned} \text{sensitivity} &= P(\hat{A} \mid A) = P(\text{model prediction} + \mid \text{observed} +) \\ P(\text{Type II error}) &= P(\hat{I} \mid A) = P(\text{false negative}) = 1 - \text{sensitivity}. \end{aligned}$$

The complementary rates are obtained from the opposite conditioning: the fraction of model-predicted hits that are observed hits $(A \mid \hat{A})$ and the fraction of model-predicted inactives $(I \mid \hat{I})$ that are observed inactives. We call these the "positive discovery rate" and "negative discovery rate". It is important to look at these conditional probabilities; a very clear example is in the analysis of gene chip microarray data where the false discovery rate is 1 minus the positive discovery rate as defined above and in Chapter 6; an excellent discussion is given by Benjamini and Hochberg (1995).

An example in the context of blood–brain barrier (BBB) predictions (see Section 6.5) is shown in Figures 8 and 9. Data from different laboratories at Lilly and from various literature sources were pooled together and molecules were assigned binary class labels, BBB+ and BBB−, depending on whether they crossed the blood–brain barrier. A random forest model, defined in Section 6.3, was trained on this data set and molecules that were not part of the training set (called "out-of-bag" in the bagging or random forest terminology) were predicted to be hits BBB+ or inactive BBB− according to a particular score/decision threshold. These predictions were evaluated and three rates were examined: sensitivity, specificity, and positive discovery rates—shown as a function of decision threshold in Figure 8, where the scores are multiplied by 10. If the goal is to obtain equal sensitivity and specificity rates (a common practice), then the optimal threshold is 0.778. Because both sensitivity and specificity are conditioned on the observed class labels, we feel it is important to include a rate that conditions on the predicted score or class label. Thus we include the positive discovery rate in our analysis.

Balancing these three rates equally yields an optimal threshold of 0.846. Both thresholds are indicated by vertical lines in Figure 8. Figure 9 shows the actual predicted scores for the molecules that do cross the blood–brain barrier (BBB+) as well as those that do not (BBB−). The false positive and false negative rates are, of course, direct consequences of which threshold is chosen. The appropriate threshold depends on the goal. For example, if the project is a neuroscience project where BBB+ is the goal, it may be that the team wants to find and reject

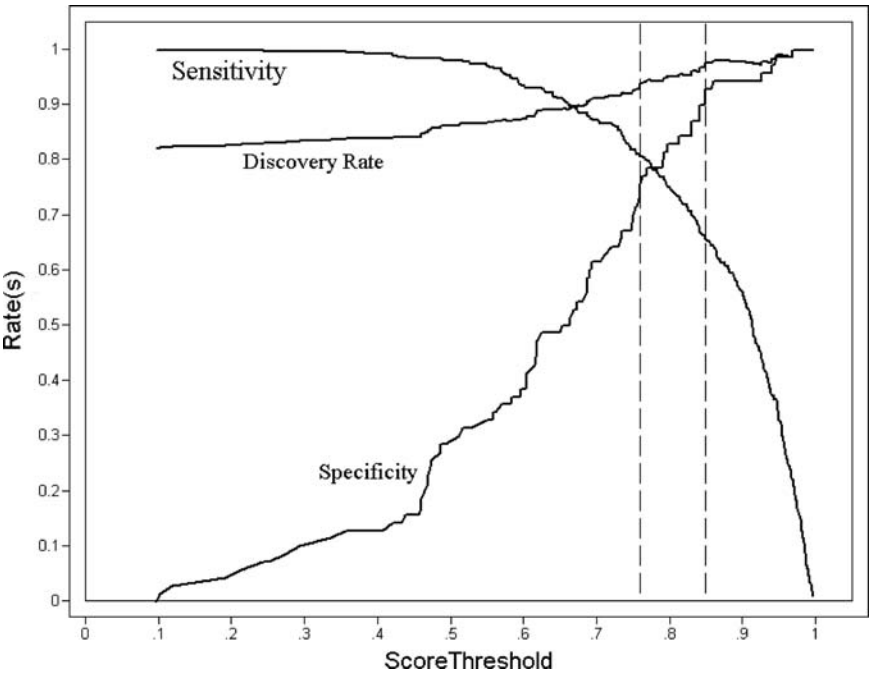



FIGURE 8. Sensitivity, specificity, and positive discovery rate as a function of decision threshold; the two reference lines correspond to two decision thresholds. The rates are estimated from predictions made for molecules not in the training set of the model.

compounds that are BBB− at an early point. Then, the goal would be to maximize sensitivity or to maximize the negative discovery rate (while realizing that going too far with this means losing a number of "good" compounds as well), and an appropriately large weight could be given, say, to specificity in computing the weighted average of the three rates to obtain an optimal threshold for that purpose.

## 6.3 Machine Learning Methods

Some of the more popular predictive modeling methods used in drug discovery include linear methods, tree-based methods, $k$-nearest neighbors, and kernel methods. In this section, a brief outline of these methods is given, together with references for reading and further details.

*Linear methods* include simple linear regression, multiple linear regression, partial least squares, logistic regression, and Fisher's linear discriminant analysis; see Hansch et al. (1962), Frank and Friedman(1993), and Hastie and Tibshirani (1996b). *Tree-based methods* are some of the most widely used methods today; see Breiman et al. (1984) and Rusinko et al. (1999). *Bagging* is a generic strategy

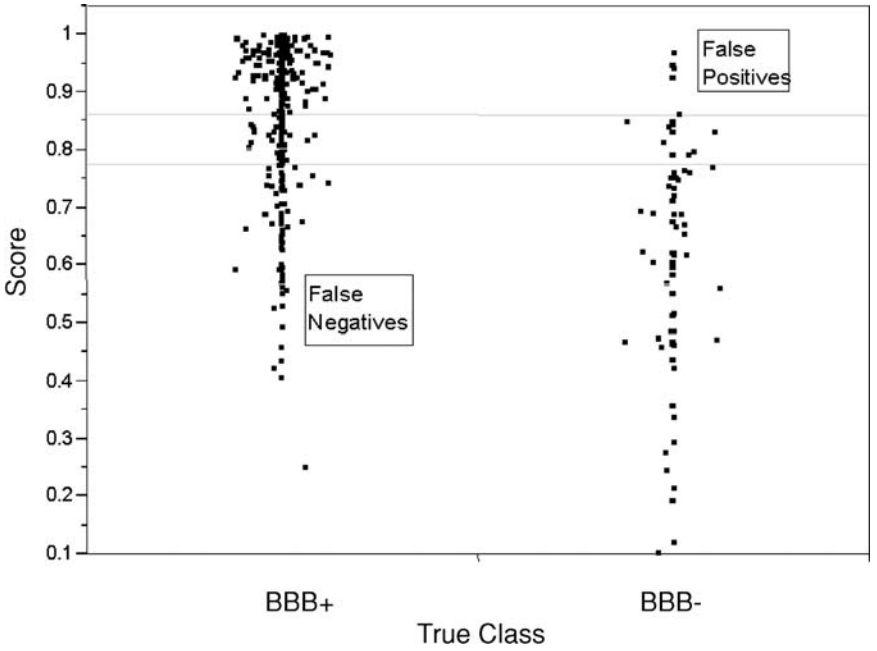



FIGURE 9. Random forest BBB predicted scores for molecules assigned as BBB+ and BBB−; horizontal reference lines correspond to two decision thresholds. All predictions (scores) are for molecules not in the training set.

that is useful in many contexts including tree-based methods. It was introduced by Breiman (1996) who motivated the strategy through the concept of unstable predictors. The bias and variance properties of aggregated predictors were further studied by Breiman (1998). *Random forests* is an improvement to the strategy of tree-based models combined with bagging. Details are given by Breiman (1999, 2001a). This is currently the top-rated algorithm in our project work at Lilly.

A simple, yet useful, and often highly accurate method is that of *K-nearest neighbors* described, for example, by Fix and Hodges (1951) and Dasarathy (1991). A notable recent advance in this method is given by Hastie and Tibshirani (1996a,b). In the case of a single descriptor, *kernel regression* and *smoothing splines* are useful methods of model fitting. However, far more general is the recent development known as *support vector machines*. This method is based on a particular hyperplane in the descriptor or property space that separates the active from the inactive compounds. This plane has the largest possible distance from any of the labeled compounds and is known as the maximum margin separating hyperplane. Support vector machines avoid overfitting by choosing the maximum margin separating the hyperplane from among the many that can separate the positive from negative examples in the feature space. Good starting points for reading about this topic are Burges (1998), Weston et al. (2002), and Vapnik (2000).

*Transduction* is another generic strategy that is an important recent advance. Standard practice in machine learning is to use *inductive learning*, that is, taking known molecules and, through training or fitting a model, generating a general understanding of the underlying relationships and then applying that general knowledge to make a prediction about a new molecule, for example, whether it will cross the blood–brain barrier. If the ultimate goal is to make predictions for a finite set of observations, then the rationale behind transduction is that the inductive learning step is not necessarily needed. Transduction skips the inductive learning step and goes directly to the prediction of the future examples. A nice heuristic explanation of this is given by Vapnik (1998, page 355). The general model that is best when applied to a universe of observations may not be the model that is best for the specific subset of observations under current scrutiny.

## 6.4 Model Selection and Assessment

Usually a variety of different models can be applied to the same data set, each model capturing part of the structural information relating explanatory variables to responses and also part of the noise. The objective of model selection may be considered, in a general sense, to be that of optimizing the quality of inference. In practice this can take several forms including discovering the "true" model, interpreting or understanding what natural process is driving a phenomenon, or simply choosing the model that gives the most accurate predictions on new data. In the QSAR drug discovery context, this latter objective is most often the appropriate one.

It is important to distinguish between algorithms and models. An algorithm creates a model, given data and tuning parameters as input. The model is a static entity. At Lilly we perform studies to select the best algorithm for a data set as well as the best model for a given algorithm, and finally to assess the error for a given model. A crucially important issue in model selection is the issue of model complexity, because training set error tends to decrease and test set error tends to increase with increasing model complexity; see, for example, Hastie et al. (2001), pages 194–199.

For the example of variable subset selection in multiple linear regression, the $R^2$ statistic increases monotonically as the number of variables added to the regression model increases, leading to the situation dubbed as *overfitting*. Various methods have been devised for avoiding an overfitted model. Some methods are simple adjustments to the familar $R^2$ statistic, such as the adjusted $R^2$ ($R^2_{adj}$) which adds a simple penalty for the number of covariates included in the model. Other popular methods include the Bayesian Information Criterion (*BIC*), the Akaike Information Criterion (*AIC*), and Mallow's $C_p$; see, for example, Burnham and Anderson (2002). In the context of high- or medium-throughput screening, when little is known about a target or an SAR, and designed experiments are not possible, there are no a priori models that can be assumed and, in any case, the key interest in early stage screening is in predictive accuracy of models rather than inference about model parameters.

TABLE 3. Size of the model space for multiple linear regression (MLR) with $h$ descriptors and for binary tree models.

| $k$ | MLR | Tree model |
|---|---|---|
| 1 | 2 | 2 |
| 2 | 4 | 9 |
| 3 | 8 | 244 |
| 4 | 16 | 238,145 |
| 5 | 32 | 283,565,205,126 |

When the model space is large, the problem becomes extreme. One solution is model averaging, in line with Breiman (1996). One good use for this approach is in recursive partitioning or tree-based models. The model space for recursive partitioning is huge. Consider the special case of binary descriptors and an algorithm that iteratively partitions data into two parts, depending on descriptor values. Once a descriptor is used to split the data, it can never be used again. Thus the model space is much smaller than when the descriptors have more than two values. For binary descriptors, the number of possible tree models $T(h)$ for a data set with $h$ descriptors can be computed from a simple recursive formula:

$$T(h) = 1 + h \cdot [T(h-1)]^2, \tag{2}$$

where $h = 0$ corresponds to the case of no descriptors where the tree model is the null model composed of the overall mean. For multiple linear regression and a simple additive model, there are $2^h$ possible models for $h$ descriptors. There is a rough analogy between the choice of parameters in the regression model and the choice of cutpoint along each descriptor in recursive partitioning. Table 3 shows the size of the model space for multiple linear regression and for binary descriptor, two-way split tree models, for up to five descriptors.

With great flexibility in model choice comes great power but also great danger of misuse. As the model space spanned by tree models is huge for $h \geq 4$, there is need for both a computationally feasible way to search the space and for some way to guard against finding spurious relationships in the data. The bagging method of Breiman (1996) was a key advance in this area.

For regression models, one metric that we use for sorting the molecules by their predicted activity, which is considered proprietary at Lilly, is similar to a weighted variant of Spearman's $\rho$. This metric, labeled $S$, ranges from $-1$ to $+1$ and compares the predicted and actual responses. The weights are higher earlier in the sorted list to emphasize that, in practice, it is the top of the sorted list that will identify the molecules selected for testing, and that accuracy farther down the list is not nearly as important. We have very little interest in accurately distinguishing the relative activity levels of molecules that are all considered inactive, but a great deal of interest in the degree to which actives will rise to the top of a sorted list of molecules. Quality assessments have been assigned to various values of $S$, but these levels in isolation are not meaningful; a very high value of $S$ (or $R^2$, or any

other metric) can easily be obtained for observations that are in the training set of a model, but does not predict how the model will perform on untested molecules. The thresholds established are based on appropriate test hold-out results, as described below. With this in mind, a value of zero is equivalent to random (the mean value resulting from scrambling the predicted responses and computing *S* many times). A value of 0.40 is considered a minimum standard for a model to be used for decision making at Lilly. Such a model would be considered weak and would not be used at all by some scientists. A model with an *S* value of 0.60 is considered a solid and useful model and an *S* value above 0.80 indicates a very good model. A difference in values of less than 0.05 is not considered to be meaningful. Thus, if one were doing model selection and two competing models were statistically significantly different but the difference in mean *S* were below 0.05, the two models would be treated as equivalent. At Lilly, we couple the concept of significant differences with the concept of meaningful differences.

## 6.5 *Example: Blood–Brain Barrier Penetration*

We examine a data set of 750 molecules with blood–brain barrier penetration measurements. An important aspect of drug design is the consideration of the potential for penetration of the blood–brain barrier by any new candidate drug molecule. Whether the goal is for the potential drug to cross or not to cross the blood–brain barrier, the ability to estimate the blood–brain ratio is an essential part of the drug design process. Determination of this aspect of a molecule is a low-throughput operation and thus having the ability to prioritize molecules in silico through the use of predictive models adds considerable value to the drug discovery process.

The penetration of a compound across the blood–brain barrier is measured experimentally as the ratio BB of the concentration of the compound in the brain to that in the blood. This ratio is thought to be related to local hydrophobicity, molecular size, lipophilicity, and molecular flexibility (Crivori et al., 2000), but no explicit mathematical relationship has been given. The 750 available results, from an in situ experiment with rats, are responses known as *Kin* values. These are intended to be related to the BB ratio of these compounds in humans. The current goal of the analysis is to select a subset of descriptors (covariates) from about 1000 possibilities and a modeling method that gives good predictive accuracy of the Kin. At Lilly, the modeling methods used do the subset selection intrinsically. In this example we compare just two methods: one is based on partial least squares (PLS) with a model-averaging strategy, and the other is the random forest algorithm of Breiman (1999).

We split the data set into two equal parts at random. We train our algorithms on one half and score the other half as test data. The idea is to study how the methods behave on new untested molecules. Whether a 50/50 split is the best choice is discussed shortly; here we consider the question of how many repetitions are needed. We started with 200 repetitions, each a random 50/50 split of the 750 data points into equal-sized training and test sets, where after each split the model

was retrained and the test hold-out set was scored. The conclusion favored the random forest model over partial least squares, and a natural question arises as to whether smaller tests would lead to the same conclusion.

## 6.6 Training and Test Set Sizes

Two key questions in model selection are what proportion of molecules to use for a training set versus a test set when doing random splits of the data, and how many different training/test splits should be analyzed to obtain reliable inferences about model performance. The number of repetitions needed is surprisingly low and often the same decisions are made whether the number of training/test splits used was 200 or 20 or 10.

Miller (2002, pages 148–150) recommended fivefold to tenfold validation, so that effectively 80% to 90% of the data should be in the training set. Another recommendation is that $n^{3/4}$ of the data should make up a training set (randomly selected) and the rest predicted as test hold-out data; see Shao (1993) for details. However, it is easy to show that use of the $n^{3/4}$ rule does not perform well in settings such as drug discovery where prediction accuracy, rather than selection of the true model, is the objective. We are sometimes better off with a model that is not the true model but a simpler model for which we can make good estimates of the parameters (leading to more accurate predicted values).

In order to choose the model that predicts most accurately for the test data, we need a new rule or a new information criterion. The usual criteria, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), Leave One Out (LOO or Qsquared), and so on, are all insufficient for our needs. This motivated Kerry Bemis to propose a new measure which he called predictive $R^2$ or $pR^2$ (described below).

Although this is an area of ongoing research, the current opinion at Lilly is that, when comparing candidate modeling methods in a model selection exercise, it is best to look at the entire learning curve (leave 90% out up to leave 10% out) and make a judgment about learning algorithms based on the performance across the whole curve. This we call a *learning curve* (but note that the phrase is used in other contexts with other meanings). Figure 10 shows the performance of the two candidate modeling methods applied to the BBB data set of Section 6.5. We generated 20 sampling runs for each level of Ptrain, where Ptrain is the proportion of the data assigned randomly to the training set, and used the two methods over a broad profile of training set sizes. The two lines connect the means of the values of $S$ obtained for the two methods at each Ptrain level. The same train/test splits were used for both the PLS and the random forest methods. Thus a paired or block analysis was done. Here, we could ask whether the random forest method is superior over all Ptrain levels and use a test such as Tukey's HSD (Honestly Significant Difference; see Tukey, 1997, Kramer, 1956). This is perhaps conservative in that we are not interested in all of the pairwise comparisons but, as we can see from simply looking at the plot, any formal comparison is going to give an unambiguous result for this data set.

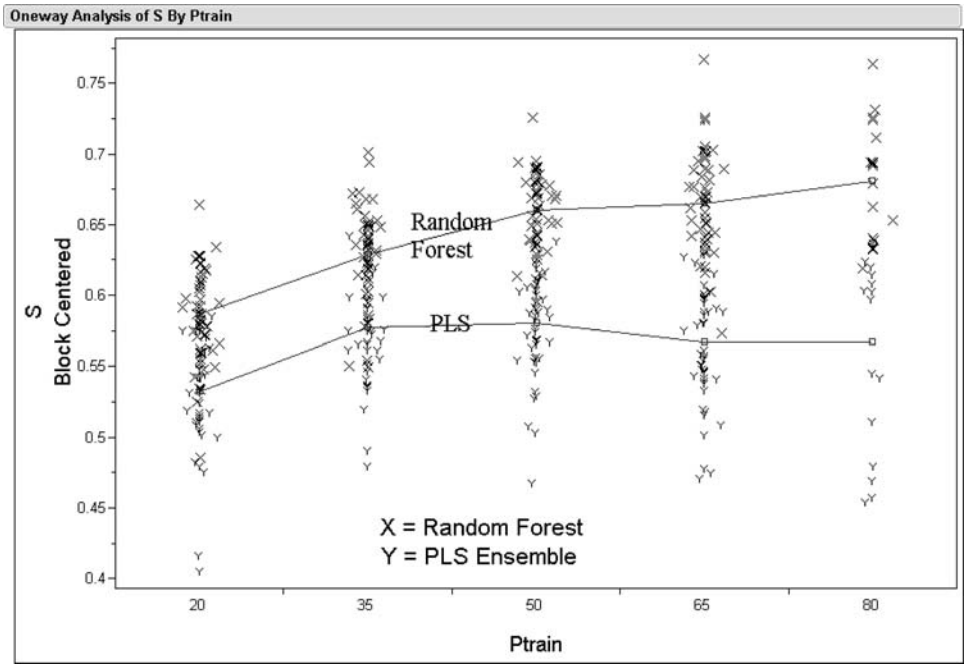



FIGURE 10. Model selection and assessment diagnostic: performance measure *S* for random forest and partial least squares (PLS) methods applied to the BBB data for various percentages of the data (Ptrain) in the training set.

There is a minimum size of training set necessary for a statistical model to be able to reveal links between vectors of descriptor values and biological activity. This has been called "statistical traction" by Young et al. (2002). Suppose a particular pharmacophoric feature is important for the binding of molecules to a receptor. Having one molecule that binds and has that feature is not sufficient for that feature to be detected as significant. Several examples of molecules that bind and contain that feature are needed before the statistical algorithm can detect it. In the model selection stage, it is possible to place a downward bias on the estimate of the predictive power of an algorithm by selecting for the training set a subset of the data that is too small. There may be a lack of "statistical traction" in the training subset that would not exist when the model is trained on all the available data. On the other hand, when the proportion of data selected for the training set is very large, and the test set is correspondingly small, it is more likely that a given test set molecule has a very similar "neighbor" in the training set and this gives an upward bias to the estimate of predictive power of the model.

Once the choice of modeling method has been made, all available data are used to train a final *scoring model* (to be used in the third step of Figure 7). Sometimes sampling issues arise here; for example, the data sets can be very large, and for classification data there is usually a huge imbalance in the number

of examples of one class compared with another. In drug hunting, there may be 400,000 examples of inactive compounds and as few as 400 active compounds. If the available modeling methods do not deal well with this situation, there may be motivation either to create a training set that alleviates such a mismatch, or to create a smaller training set to reduce the computational burden. Either of these issues may or may not be related to the problem of model selection. One strategy for selecting a subset of available data for training a model is as follows.

1. Select all the active compounds.
2. Select a small subset of the inactive compounds whose nearest neighbor among the active compounds is a relatively short distance (by some distance measure such as those of Section 4.5). The motivation here is to preserve the boundary between classes.
3. From the remaining inactive compounds, select a maximally diverse subset (as described in Section 5). This augments the space beyond the boundary with an optimal exploration of the chemical space represented by inactives.

At Lilly we have focused on predictive accuracy in most of our project work. Predictive accuracy and interpretability tend to be inversely proportional. An active area of research at Lilly is an investigation of the question of ways in which the model can help us design a better molecule. This may involve interpretation, and there are excellent tools that can be used for this, such as partial dependence plots. It can also be approached through virtual screening—a scientist proposes a scaffold or series and the model provides an evaluation of the prospects of that idea.

## 6.7 *The Predictive $R^2$ of Bemis*

In the area of linear models, Bemis has proposed a "predictive $R^2$" or $pR^2$. Until 2004 this criterion was treated as a trade secret at Lilly. The $pR^2$ does not involve training/test split cross-validation, but rather uses an information-theoretic criterion motivated by ideas of Shi and Tsai (2002). For a model with $h$ parameters,

$$pR_h^2 = 1 - \exp\left[\frac{RIC_h}{k-h-1} - \frac{RIC_0}{k-1}\right],$$

where $RIC_0$ corresponds to the Shi and Tsai RIC (residual information criterion) for the null model. To give more clarity, we give an alternative notation of the $pR^2$, building up from the familiar $R^2$ to the adjusted $R^2$ and finally to the $pR^2$. For a linear model with $h$ parameters:

$$R_h^2 = 1 - \frac{SSE_h}{SSE_0} = 1 - \frac{SSE_h/k}{SSE_0/k} = 1 - \frac{\widehat{\sigma}_h^2}{\widehat{\sigma}_0^2},$$

$$adjR_h^2 = 1 - \frac{SSE_h/(k-h-1)}{SSE_0/(k-1)} = 1 - \frac{\tilde{\sigma}_h^2}{\tilde{\sigma}_0^2} = 1 - \exp\left[\log\left(\tilde{\sigma}_h^2\right) - \log\left(\tilde{\sigma}_0^2\right)\right],$$

$$pR^2 = 1 - \exp\left[\left\{\log\left(\tilde{\sigma}_h^2\right) + bias_h\right\} - \left\{\log\left(\tilde{\sigma}_0^2\right) + bias_0\right\}\right],$$

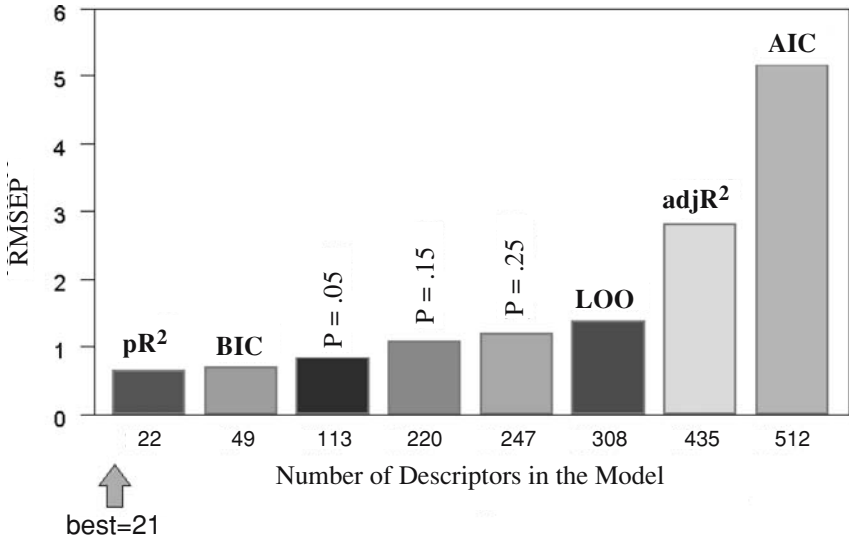



FIGURE 11. Performance of competing criteria: the number of descriptors in the model, for various criteria versus the root mean squared prediction error (RMSEP) in forward selection. (Reproduced with permission from the author.)

where

$$bias_h = \frac{h+1}{k-h-1}[\log(k) - 1] + \frac{4}{(k-h-1)(k-h-3)}.$$

Figures 11 and 12 illustrate the performance of the $pR^2$ compared with several of the currently popular criteria on a specific data set resulting from one of the drug hunting projects at Eli Lilly. This data set has IC50 values for 1289 molecules. There were 2317 descriptors (or covariates) and a multiple linear regression model was used with forward variable selection; the linear model was trained on half the data (selected at random) and evaluated on the other (hold-out) half. The root mean squared error of prediction (RMSE) for the test hold-out set is minimized when the model has 21 parameters. Figure 11 shows the model size chosen by several criteria applied to the training set in a forward selection; for example, the $pR^2$ chose 22 descriptors, the Bayesian Information Criterion chose 49, Leave One Out cross-validation chose 308, the adjusted $R^2$ chose 435, and the Akaike Information Criterion chose 512 descriptors in the model. Although the $pR^2$ criterion selected considerably fewer descriptors than the other methods, it had the best prediction performance. Also, only $pR^2$ and BIC had better prediction on the test data set than the null model.

## *6.8 Common Errors*

Predictive modeling, statistical modeling, and machine learning are very open areas in the sense that the barrier to admission is very low. All that is needed to

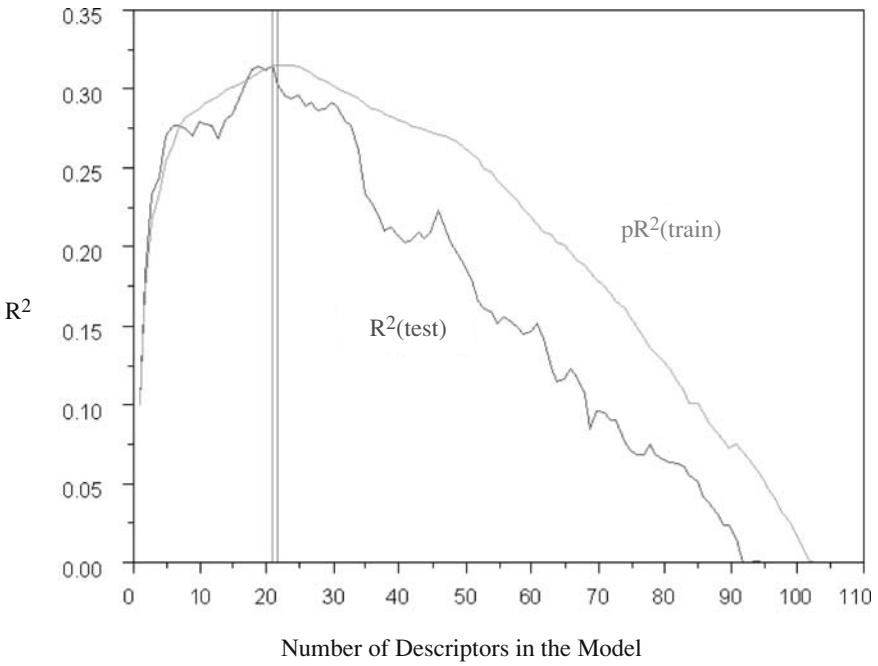



FIGURE 12. The Bemis $pR^2$ for an example data set. The "true" observed $R^2$ based on the test set, and the $pR^2$ estimated only from the training set. (Reproduced with permission from the author.)

start experimenting in this area is a PC and a data analysis package. Below is a list of the most frequent errors as they occur in this field.

1. Belief that a very small $p$-value for a predictor (for example, a biomarker) is more likely to occur with high predictive accuracy. The multiple testing problem must not be ignored, and the false discovery rate (FDR) controlled; see also Chapter 6.
2. Failure to pre-process and clean the data. Sometimes even data with missing values are jammed through a learning algorithm with little thought.
3. Use of an unsupervised algorithm to do the job of a supervised algorithm. For example, a cluster analysis or self-organizing map is used in combination with a post hoc analysis to do prediction.
4. Failure to evaluate a method on test data.
5. Test data set too small, with these consequences:
   a. Prediction error cannot be accurately estimated on each hold-out part.
   b. The test sample and the training sample are likely to be similar in their descriptor values.
6. "Cheating" in model assessment: using the whole training set to select descriptors, and then splitting the data into train and test sets, and training the model on the descriptor set selected from the whole training set.

7. Comparison of methods with the same type of feature selection forced on all methods, rather than letting each method do what it does best.
8. Confusion of model selection with model assessment. If one chooses the model with the lowest cross-validated error among competing models, that error is not a valid estimate of the prediction error of that model (selection bias).

## 7 Iterative Medium-Throughput Screening

Researchers who use high-throughput screening (HTS) methods are troubled with many obstacles such as poor data quality, misleading false-positive and false-negative information, and the need to confirm and expand the SAR of the identified lead candidates. Additionally, HTS strategies lead to the large-scale consumption of valuable resources such as proteins and chemicals from the inventory, and may not be applicable to all targets (Major, 1999). The problem is fueled, in particular, by the prospects of expanding universes of targets—an increase by a factor of 10 is expected (Drews. 2000)—that will lead to an explosion of costs. As a consequence, there is a need not only to increase the scope of screening, but also the efficiency of each screening experiment. Hybrid screening strategies have been suggested that unite in silico and in vitro screening in one integrated process.

Iterative medium-throughput screening (MTS) starts with a small (200 to 20,000) and "diverse" subset of compounds. This initial sample is subjected to a primary screening where the main objective is to gather SAR data for predictive model building. This is a key distinction from the older paradigm where the primary objective is to obtain an initial set of screening hits, and any subsequent model building is an added bonus. Based on this first SAR, the corporate inventory is screened in silico in order to identify a further, more focused set of compounds, the focused library, for a second round of MTS. Several cycles of testing–analyzing–testing can be applied aimed at either refining the SAR model(s) or the identification of more active compounds. Abt et al. (2001) studied the influence of the size of the focused sample and the number of cycles on the effectiveness of the computational approaches.

One factor that plays a role in the decision on how compounds are chosen in a given cycle is the stage of a project. An early stage project may require more diversity to be built into the selection process, whereas a later stage lead optimization effort would draw much more heavily on predictive modeling and expert judgement.

*Active learning* is a strategy that is iterative and where the selection of compounds to test in the next iteration is based on all the currently available data (see, for example, Warmuth et al., 2003 and Campbell et al., 2000). A key distinction between active learning and other modeling strategies is that, in active learning, the primary objective for selecting the next batch of compounds for testing is to improve the model optimally, whereas in drug screening programs the primary goal is to find as many potent (and usually novel) compounds as possible. This distinction and the consequent effects are dramatic. In active learning, the most

interesting compounds are the ones for which the model has had difficulty in the assignment of a clear classification whereas, in a typical drug hunting program, the most interesting compounds are the ones that are scored the most unambiguously as active. This has downstream implications on what will come out of future iterations of screening. The traditional business-driven approach will find good compounds faster, but the active learning approach will generate better models faster, and eventually lead to better exploration of chemical space, resulting in finding the best compounds.

There is also an analogy with ancillary efforts such as toxicity testing. A drug hunting project tends to focus on finding compounds with potency and selectivity for the target of interest. When interesting compounds are found, they are submitted for toxicity testing so that a small set of structurally related compounds is tested for the toxic endpoint of interest. This places a handicap on any toxicity modeling effort. If the goal is to develop a good toxicity model (which would reduce the need for animal testing and reduce cycle time in the project), then compounds that are not interesting from a potency standpoint would need to be tested for toxicity. This would mean that the interesting potent compounds must wait their turn due to limited capacity in toxicity testing. The long view, both in terms of active learning for potency and for toxicity, might be to strike some balance between immediate and future gains.

## 8 Virtual Screening and Synthesis

*Virtual screening* is a simple concept, arising from the need to break out from the confines of the currently available set of in-house chemical libraries. It is a simple matter to construct representations of molecules using computers, and this can be done in a combinatorial manner. Usually one or more "scaffolds" are chosen—these are the "backbone" of the molecule. Then a number of "substituents" are chosen; these can be thought of as the "appendages" that gets attached to the backbone at various (preselected) locations. All (reasonable) combinations of scaffolds and substituents can be made in silico and these structures form a virtual library. The library may contain millions of molecules but it is more typical to see something of the order of 500,000 structures. This is because most virtual screening efforts are knowledge driven; something is known about the SAR before the virtual screen is attempted. Most of the molecules in the virtual library will not exist in the corporate molecular stores. This virtual library is then the subject of a modeling effort whereby the virtual library is prioritized and rank ordered, with the most promising structures at the top of the list. The biological screening is done virtually through the use of the predictive models applied to the virtual library.

Some of the high-ranking structures may be very similar to structures that have already been tested. These are removed from the list using molecular diversity methods such as a Leader algorithm (Hartigan, 1975). In this context, the Leader algorithm is not providing a cluster analysis, but simply a post-processing of a rank

ordered list. From what is left, a relatively small number of molecules are then synthesized and tested for biological activity. Because this is a relatively expensive part of the process, it is usually important that some knowledge about the SAR has been gained before the virtual screening is done. That knowledge could come from literature sources or from prior early-stage screening.

## 9 Pooling

Pooling strategies can take numerous forms, as discussed in Chapter 3. In the drug hunting screening context, chemical compounds can be pooled. Ten compounds may be pooled together in a well and tested as a mixture. If the mixture is potent, the individual components can then be tested. If the mixture shows no potency, it might be assumed that the individual components are each inactive. This assumption may sometimes be incorrect, as compounds may exert an antagonistic (or conversely, synergistic) effect on each other. For the use of orthogonal arrays in the design of a pooling study see Phatarfod and Sudbury (1994; and also Dorfman, 1943).

The design and deconvolution of pools in drug discovery screening has been approached in different ways by a number of companies. In a highly specialized experiment at Merck, Rohrer et al. (1998) pooled a staggering 2660 compounds per well. The deconvolution of these results was done using chemical technology rather than the informatics approach one might use following Phatarfod and Sudbury (1994).

An interesting informatics strategy involves pooling covariates in a variable subset selection context. Suppose one has a data set with hundreds of thousands of covariates (descriptors), as happens in the drug discovery setting, and perhaps one does not have a data analysis package capable of handling so many columns of data. If the covariates are sparse binary, meaning that each column is mostly zeros with a few ones (a typical scenario), one strategy for data reduction is to pool columns together. One could take batches of, say, 100 columns and simply add them, creating a "pooled covariate." This data set is now 100-fold smaller, and a forward selection method might be run to fit a model on the reduced data set. Variables selected by such a procedure can then be collected and the individual covariates unpacked and a second stage of variable selection performed on this reduced data set.

## 10 Expectations for Discovery of Rare Events

The hit rate within a set of molecules selected by a virtual screen is primarily determined by two parameters: the unknown proportion of $p$ hits that exist in the set of molecules scored and the false positive error rate ($\alpha$) of the classifier used for virtual screening. To a large extent, the statistics of rare events (true hits within a large compound collection) leads to some initially counterintuitive results in the magnitude of a hit rate within a set of molecules selected by a model.

Most pharmaceutical companies expect to see hit rates in the 0.1% to 1% range for a high-throughput screen. In the virtual screening context, when the hits are a rare event (of the order of 0.1%) even very good predictive models cannot be expected to lead to arbitrarily high hit rates for the molecules selected. It is quite likely that marginal to good virtual screen models will result in no hits identified in a subset of molecules selected by virtual screening.

The virtual screen can be considered as a classifier that makes a prediction about whether a molecule is likely to be active or inactive in a biochemical assay. It can be constructed from training data (for example, a QSAR model) or constructed from a model of a binding site. For a given molecule in a virtual library, let the null hypothesis be that the molecule is not a hit. Then, using the notation of Section 6.2,

$$P(\mathrm{A}) = p, \qquad P(\hat{\mathrm{A}} \mid \mathrm{A}) = 1 - \beta, \qquad P(\hat{\mathrm{A}} \mid \mathrm{I}) = \alpha,$$

$$P(\mathrm{A} \mid \hat{\mathrm{A}}) = \frac{P(\hat{\mathrm{A}} \mid \mathrm{A})P(\mathrm{A})}{P(\hat{\mathrm{A}} \mid \mathrm{A})P(\mathrm{A}) + P(\hat{\mathrm{A}} \mid \mathrm{I})P(\mathrm{I})} \tag{3}$$

$$= \frac{(1-\beta)p}{(1-\beta)p + \alpha(1-p)}. \tag{4}$$

Equation (3), which is an application of Bayes theorem, is referred to as the "Positive Predictive Value." The parameter $p$ is unknown but believed to be very small ($<0.01$) for large virtual libraries. $1 - \beta$ is the power (or 1 – type II error, where $\beta$ is the false negative error rate) and $\alpha$ is the type I error, also called the "size" of a test in the hypothesis testing context, or the false positive error rate. The last equation defines the probability that a molecule is determined to be a hit in a biochemical assay given that the virtual screen predicts the molecule to be a hit. This probability is of great interest because it is valuable to have an estimate of the hit rate one can expect for a subset of molecules that are selected by a virtual screen.

The values of parameters $p$, $\alpha$, and $\beta$ can be varied to observe the effect on equation (3). It is straightforward to verify that the "power" of the classifier $(1 - \beta)$ has relatively little effect on the hit rate observed in the subset of molecules selected by a virtual screen. The influence of power is greatly reduced as the probability of a hit existing in the set of compounds being scored decreases (the low prevalence effect) and, for rare events, the relative importance of $\alpha$ is greatly intensified. Even for less rare events, say a hit rate of 10% (disturbingly high in drug discovery, suggesting nonspecificity in the assay), the effect of $\alpha$ dominates.

## 11 Drugability of Molecules: ADME, Solubility, Toxicity

The word *drugability* is often used to cover all aspects of a molecule beyond initial potency. A potential drug compound must overcome many challenges in order to be a successful therapeutic. Critical components of drug design include absorption, permeation, distribution, metabolism, stability, specificity (does it do more than

intended?), and toxicity (related but not identical to specificity). In this section, some of these issues are discussed in more detail.

## *11.1 ADME*

Many of the compounds entering clinical trials are discontinued, often due to issues directly related to *ADME*: absorption, distribution, metabolism, and elimination/excretion of a drug. *Absorption* is of paramount importance, being the extent to which an intact drug is absorbed from the gut lumen into the portal circulation. *Distribution* is important because the drug will not work if it is not transported to the intended site. A compound may have potent effects in vitro screens involving cells or enzymes, but in a living organism the compound may have no effect because of a distribution problem. This can be due to a number of things; for example, the compound may bind so tightly to proteins in the bloodstream that it does not leave the bloodstream until it is eliminated by the liver. The opposite extreme can be a problem as well, because proteins in the blood can be important as transport mechanisms. In addition, the unbound drug may penetrate the wall of the blood vessel so that a certain amount of protein binding is desirable. Most pharmaceutical companies have models that predict the protein binding affinity of compounds. Distribution is only one problem that can confound an SAR effort when transitioning from in vitro to in vivo screens.

Two endpoints important to distribution are *oral bioavailability* and *first pass clearance*; see Birkett (1990, 1991). Oral bioavailability is particularly important because a drug that has, say, only 10% oral bioavailability would require a 10-fold higher dose when given orally as compared with being given intravenously. Orally administered drugs, after absorption through the gut lumen into the portal circulation, must then pass through the liver before reaching the systemic circulation. Pre-systemic or first pass extraction refers to the removal of drugs during this first pass through the liver. *First pass clearance* is the extent to which a drug is removed by the liver during its first passage from the portal blood on its way to the systemic circulation. Oral *bioavailability* is the fraction of the dose that reaches the systemic circulation as intact drug. It is apparent that this will depend both on how well the drug is absorbed and how much escapes being removed by the liver. In fact, the simple equation for bioavailability is

$$Ba = \text{fraction absorbed} \times (1 - \text{extraction ratio}),$$

where the extraction ratio is the proportion removed by the liver. Thus if drug $\mathcal{A}$ has 80% absorption and 75% extraction ratio, then the bioavailability of $\mathcal{A}$ is 20%. The 20% alone does not tell us anything about the metabolism or the absorption of the drug.

Because there are many ways to achieve a given level of bioavailability, it makes sense to consider using a compartmental model to predict bioavailability rather than simply training a model on a set of bioavailability results. The role of metabolism tends to dominate most often and variability in drug response is greatly influenced by this. Drugs that are efficiently eliminated by the liver often have high variability in the plasma levels both within and between individuals

because, in that case, slight changes to the extraction ratio can cause large changes to the resulting bioavailability.

Treated as a special case of distribution is the ability of a molecule to cross the blood–brain barrier (BBB). This fact is important to know, both for central nervous system (CNS) drugs and for drugs that do not target the central nervous system. There has been a flurry of research attempts to model and/or predict the BBB propensity of molecules. Many of these efforts are statistically destitute; for example, a research group may examine only a set of molecules that do cross the BBB. Proper inference must involve examples of compounds that do not cross the BBB as well as compounds that do and this falls in the domain of predictive modeling and machine learning (see Section 6). The BBB is formed by the highly selective capillaries of the central nervous system. Passage of drugs through the BBB may occur by passive diffusion or from various specific uptake mechanisms, many of which are there to supply nutrients to the brain. There are also mechanisms for transporting substances out of the brain. P-glycoprotein (or Pgp) is an efflux pump that removes many drug compounds from the brain. Thus BBB transport is a complex phenomenon and modeling this is a challenging and ongoing research topic in most pharmaceutical companies.

*Metabolism* is another critically important aspect for determining the fate of a drug. If a drug is metabolized quickly, it may be excreted in the urine before it has a chance to reach the intended site, but the full story is much more complicated than this. Most successful drugs are lipid-soluble and are reabsorbed from the kidney back into the bloodstream. These compounds undergo metabolism, which is a way for the body to break down and ultimately eliminate a substance. The liver uses a number of different enzymes to break a compound down into smaller parts, called metabolites. A metabolite may either be pharmacologically similar to the parent compound or harmless, but not pharmacologically active, or may possess life-threatening toxicity. Thus it is essential to know into which of these categories a drug falls and it is desirable to control this aspect in a favorable way. Ideally a compound would metabolize at a moderate rate, neither too slowly nor too quickly. Because humans are genetically diverse, the same compound will be metabolized differently in different people. All of these issues are interdependent and are illustrated in the following examples.

A major group of enzymes, not just in the liver but also in the intestines, lung, kidneys, and brain, is known as the Cytochrome P-450 isoenzymes, often abbreviated as *CYP450*. Some drugs interact with these enzymes. A drug with a high affinity for an enzyme will slow the metabolism of any low-affinity drug; for example, grapefruit juice inhibits a number of CYP450s which results in higher than expected levels in the body of the drugs that are metabolized by those CYP450s.

The inhibition of CYP450 isoenzymes by grapefruit juice lasts about 24 hours and occurs in all forms of the juice—fresh fruit and fresh and frozen juice. There is the potential for dangerous arrhythmias for patients taking cisapride, astemizole, and terfenadine. Other substances may induce the opposite effect, that is, upregulate the levels of the enzymes and result in faster metabolism, for example, smoking and the ingestion of charbroiled meats may induce isoenzymes, resulting in increased clearance of drugs (such as theo-phylline). The herb known as St. John's wort

causes an increase in the Cytochrome P450 enzymes, especially CYP 3A4, which are responsible for the metabolism and elimination of many drugs. This is why the birth control pill is rendered less effective by St. John's Wort. But in addition to being an inducer of CYP 3A4, St. John's Wort is also an inhibitor of CYP 2D6. Hence patients taking St. John's wort are likely to experience an increase in blood levels of therapeutic drugs that are metabolized by the 2D6 family (this includes beta blockers, antidepressants, antipsychotics, cough suppressants, codeine, and others) as well as a decrease in blood levels of drugs that are metabolized by the 3A4 family of CYP 450s (which includes antibiotics, HIV protease inhibitors, antihistamines, calcium channel blockers, and others).

Just two decades ago, the FDA was uninvolved in issues regarding CYP450 metabolism, but currently there are stringent guidelines that must be met to ensure that the metabolic fate of a drug is under control. There are genetic polymorphisms in some of the genes expressing CYP450 subfamilies. For example, 5 to 10 percent of Caucasians have polymorphic forms of the 2D6 subfamily; such individuals are called "slow metabolizers." There is a large list of drugs metabolized by 2D6 that can pose a risk to slow metabolizers and dosing must be done carefully.

## *11.2 Solubility*

Solubility plays a critical role in the absorption of a drug. A compound with poor solubility may not achieve high enough levels in the stomach and intestine to be absorbed well. However, it is generally true that highly soluble compounds lack sufficient lipophilicity to cross the blood–brain barrier and so, if the compound is an intended CNS drug, a balance must be maintained; see Amidon et al. (1995).

## *11.3 Toxicity*

Toxicity is often related to ADME; for example, when a compound cannot be broken down and eliminated by the body it builds up toxic levels in the system. Some other toxicity issues that have recently received heightened attention are discussed below. Phospholipidosis (an adaptive storage response to drug administration) and cardiomyopathy (a pathologic condition of the heart muscle) have been reasons for the recent FDA withdrawal of drugs. Another issue concerns the need for additional assurance of the absence of any potential for QT prolongation (an effect on electrical impulse conduction in the heart). Many classes of drugs induce QT prolongation, including antihistamines, antibiotics, antipsychotics, and macrolides. QT prolongation can lead to sudden death. At least four drugs have been taken off the market due to QT prolongation alone: Terfenadine, Sertindole, Astemizole, and Grepafloxacin. Acquired long QT syndrome (LQTS) occurs as a side effect of blockade of cardiac HERG K+ channels by commonly used medications. These issues have inspired modeling efforts aimed at predicting these effects and using those predictions to filter compounds in the early screening stages. Because these models are less than perfect, false negative and false positive rates are an issue.

A common approach to lead optimization is now parallel optimization, which is discussed in the following section. This is an extremely challenging undertaking because it requires the simultaneous control of several medicinal chemistry components. Furthermore, many of these components are not independent and, in fact, may even be negatively correlated. To do parallel optimization, it is also necessary to generate drugability related information in the early stages of lead optimization.

## 12 Multi-Objective Optimization Methods

Decisions in drug discovery are almost always multidimensional. Numerous criteria must be managed in order to develop a successful drug: potency, selectivity, toxicity, and ADME characteristics, and these tend to have conflicting trends so that difficult decisions are forced on scientists. For example, Zyprexa is an excellent antipsychotic drug but it causes weight gain in most people, a side effect of almost all antipsychotic drugs. Why this happens is still being investigated and there are at least six different hypotheses given in the literature. It is possible to modify an antipsychotic drug so that it does not produce weight gain, but such modifications may reduce the potency of the drug or introduce other side effects which may be even worse. A common side effect, for example, of many antipsychotic drugs is "extrapyramidal side effects" (EPS) which produce symptoms such as tremors, rigidity, and slowness of movement. These are deemed by most to be worse than weight gain. Less clear-cut trade-offs might involve the propensity for a molecule to cross the blood–brain barrier versus the therapeutic effect desired. For example, Benadryl is still a popular drug because, in spite of its tendency to induce a feeling of somnolescence, it is an extremely potent histamine (H1) blocker. Specificity is a problem faced by virtually every project team in drug discovery. Potency is desired at one receptor but not at another.

The old paradigm in drug discovery, which might be labeled "sequential search," generally fails. With this paradigm, one would optimize each objective independently and in succession. Finding a lead compound corresponds to searching on one landscape. Optimizing the lead corresponds to searching additional landscapes starting with the results of searching the first. With more than two objectives, the likelihood of failure increases exponentially. What is needed is a holistic approach with a mathematical framework for considering trade-offs between objectives. A variety of algorithms exists for finding the best possible trade-offs; these are used surprisingly seldom despite their utility.

One strategy involves restricting a search to only those solutions that are *Pareto optimal*. A solution is Pareto optimal if there is no other solution that is better under one criterion without being worse for the other criteria. It is often true that not every response has the same importance; for example, avoiding EPS symptoms might be 50-fold more important to a team than avoiding weight gain. Although Pareto optimality provides more than one solution, it does not allow different weightings on different criteria, as this is difficult to manage with

more than two dimensions. Another useful approach is Derringer's desirability function which does allow weights to be assigned to each criterion (Derringer, 1980). The desirability function involves transformation of each criterion to a desirability value $d$, where $0 \leq d \leq 1$. The transformation is done in such a way that the value of $d$ increases as the "desirability" of the corresponding criterion increases. This transformation may be linear, quadratic, step function, and so on. In the terminology of decision theory, these are monotonic utility functions. The individual desirabilities are then combined using a geometric mean, which is an overall assessment of the desirability of the combined response levels. It can be a weighted mean where the weights reflect relative importance of the criteria.

## 13 Discussion

Drug discovery is a challenging endeavor that involves many disciplines in the life sciences and informatics. There are a great many interesting and diverse problems that need to be solved. This chapter has given an overview of a number of them while omitting many others. Areas that are increasing in research intensity include the areas of genomics, gene chip microarrays (see Chapter 5), proteomics, metabolomics, and other technologies that involve spectral analysis. There are a host of interesting and challenging problems in these areas and, currently, there is great interest in merging these disciplines with the cheminformatics-related disciplines that have been the focus of this chapter.

The future will see dramatic changes in drug discovery and development processes. Within the next decade, researchers will almost certainly find most human genes and their locations. Explorations into the function of each one is a major challenge extending far into the next century and will shed light on how faulty genes play a role in disease causation. With this knowledge, commercial efforts will shift towards developing a new generation of therapeutics based on genes. Drug design will be revolutionized as researchers create new classes of medicines based on a reasoned approach using gene sequence and protein structure information rather than the traditional trial-and-error method. The drugs, targeted to specific sites in the body, will not have the side effects prevalent in many of today's medicines. Over 150 clinical gene therapy trials are now in progress in the United States, most for different kinds of cancers.

The road map of human biology generated by the human genome project will supply an enormous store of genes for studying, and ultimately curing, the ills that beset us. As the factors underlying the maladies of the human condition slowly come to light, the challenge will be to use the information effectively and responsibly.

*Acknowledgments.* I would like to thank Rick Higgs, Kerry Bemis, Bruno Boulanger, and Philip Iversen for fruitful discussions.

## Appendix
## Tools of the Trade

Robots: Used in a number of processes: screening compounds for biological activity, inoculating microbial cultures, and filling compound libraries.

High-throughput screening (HTS): Technology where robotics is used to test many compounds rapidly in an effort to identify novel inhibitors of receptors or enzymes. Usually 100,000 to 200,000 compounds are screened.

Medium-throughput screening (MTS): Similar to HTS but with only modest throughput requirements which implies more careful usage of robotics and higher quality of data. Typical MTS throughput is 1000 to 10000 compounds.

Combinatorial chemistry (Combi Chem): Used to make thousands of variants of a compound. Consider a compound with a six-membered aromatic ring and a chlorine atom attached at a certain position. One might change the location of the chlorine to any of the other five positions, or change the chlorine to a fluorine or a bromine, and/or make the same changes at all the other five positions. To enumerate all the possible combinations is to make a combinatorial library.

Genomic information: Used to identify possible protein therapies and targets, to develop biomarkers, and to understand more deeply how a given compound interacts with a complex living system.

X-ray crystallography, nuclear magnetic resonance (NMR): Used in exploring the physical properties/shape of a molecule and/or a receptor target. If the structure of the target is known, docking studies can be done to assess how well a molecule may "fit" in one of the receptor's binding sites.

Bioinformatics tools: Used to search enormous volumes of biological information, for instance, to find the best genomic match of a nucleotide sequence or learn the chromosomal location and disease linked to a particular gene. We may know that a compound evokes a biological response but with genomics and bioinformatics tools we can examine which proteins are affected by the compound.

Cheminformatics tools: Used to explore the relationship between the structure of a compound and the biological response it evokes (the SAR), with a view toward predicting what will happen with new, as yet untested compounds. Also used to model the docking of a small molecule (or ligand) to a protein or receptor.

## *References*

Abt, M., Lim, Y., Sacks, J., Xie, M., and Young, S. S. (2001). A sequential approach for identifying lead compounds in large chemical databases. *Journal of Biomolecular Screening*, **16**, 154–168.

Amidon, G., Lennernäs, H., Shah, V., and Crison, J. (1995). A theoretical basis for a biopharmaceutic drug classification: The correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharmaceutical Research*, **12**, 413–420.

Bejamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289–300.

Birkett, D. J. (1990). How drugs are cleared by the liver. *Australian Prescriber*, **13**, 88–89.

Birkett, D. J. (1991). Bioavailability and first pass clearance. *Australian Prescriber*, **14**, 14–16.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123–140.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, **26**, 801–849.

Breiman, L. (1999). Random forests, random features. *Technical Report*, University of California, Berkeley.

Breiman, L. (2001a). Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, **16**, 199–231.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. CRC Press, New York.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*. Springer-Verlag, New York.

Campbell, C., Christianini, N., and Smola, A. (2000). Query learning with large margin classifiers. *Proceedings of ICML2000*, 8.

Comprehensive Medicinal Chemistry (2003). MDL Informations Systems, California.

Cook, R. D. and Nachtsheim, C. J. (1982). Model robust, linear-optimal design: A review. *Technometrics*, **24**, 49–54.

Crivori, P., Cruciani, G., Carrupt, P., and Testa, B. (2000). Predicting blood-brain barrier permeation from three-dimensional molecular structure. *Journal of Medicinal Chemistry*, **43**, 2204–2216.

Crum-Brown, A. and Fraser, T. R. (1869). On the connection between chemical constitution and physiological action. Part I. On the physiological action of the salts of the ammonium bases derived from strychnine, brucia, thebaia, codeia, morphia and nicotia. Part II. On the physiological action of the ammonium bases derived from atropia and conia. *Transactions of the Royal Society of Edinburgh*, **25**, 151–203; 693–739.

Cummins, D. J., Andrews, C. W., Bentley, J. A., and Cory, M. (1996). Molecular diversity in chemical databases: Comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *Journal of Chemical Information and Computer Sciences*, **36**, 750–763.

Dasarathy, B. (1991). *Nearest Neighbor Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA.

Derringer, G. and Suich, R. (1980). Simultaneous optimization of several response variables. *Journal of Quality Technology*, **12**, 214–219.

Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Science*, **14**, 436–440.

Drews, J. (2000). Drug discovery: A historical perspective. *Science*, **287**, 1960–1964.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77–87.

Engels, M. F. M., Thielemans, T., Verbinnen, D., Tollenaere, J. P., and Verbeeck, R. (2000). Cerberus: A system supporting the sequential screening process. *Journal of Chemical Information and Computer Sciences*, **40**, 241–245.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Technical Report 4*, U.S. Air Force, School of Aviation Medicine, Texas.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.

Goldberg, J. and Wittes, J. (1978). The estimation of false negatives in medical screening. *Biometrics*, **34**, 77–86.

Hansch, C., Maolney, P. P., Fujita, T., and Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*, **194**, 178–180.

Hartigan, J. (1975). *Clustering Algorithms*. John Wiley and Sons, New York.

Hastie, T. and Tibshirani, R. (1996a). Discriminant adaptive nearest-neighbor classification. *IEEE Pattern Recognition and Machine Intelligence*, **18**, 607–616.

Hastie, T. and Tibshirani, R. (1996b). Discriminant adaptive nearest neighbor classification and regression. In *Advances in Neural Information Processing Systems*. Editors: D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, volume 8, pages 409–415, MIT Press, Cambridge.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Hawkins, D. M., Basak, S. C., and Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, **43**, 579–586.

Higgs, R., Bemis, K., Watson, I., and Wikel, J. (1997). Experimental designs for selecting molecules from large chemical databases. *Journal of Chemical Information and Computer Sciences*, **37**, 861–870.

JMP (2003), Version 4.0.4. SAS Institute, North Carolina.

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990). Minimax and maximum distance designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.

Kennard, R. and Stone, L. (1969). Computer aided design of experiments. *Technometrics*, **11**, 137–148.

Kramer, C. Y. (1956). Extensions of multiple range tests to group means with unequal numbers of replications. *Biometrics*, **12**, 309–310.

Leach, A. R. and Gillet, V. J. (2003). *Introduction to Chemoinformatics*. Kluwer Academic, Boston.

Maccs Drug Data Report (2003). MDL Informations Systems, California.

Major, J. (1999). What is the future of high-throughput screening? *Journal of Biomolecular Screening*, **4**, 119–125.

Miller, A. J. (2002). *Subset Selection in Regression*, second edition. Chapman & Hall/CRC, New York.

Phatarfod, R. M. and Sudbury, A. (1994). The use of a square array scheme in blood testing. *Statistics in Medicine*, **13**, 2337–2343.

Rohrer, S. P., Birzin, E., Mosley, R., Berk, S. C., Hutchins, S., Shen, D., Xiong, Y., Hayes, E., Parmar, R., Foor, R., Mitra, S., Degrado, S., Shu, M., Klopp, J., Cai, S. J., Blake, A., Chan, W. W. S., Pasternak, A., Yang, L., Patchett, A., Smith, R., Chapman, K., and Schaeffer, J. (1998). Rapid identification of subtype-selective agonists of the somatostatin receptor through combinatorial chemistry. *Science*, **282**, 737–740.

Rusinko, A., III, Farmen, M. W., Lambert, C. G., Brown, P. L., and Young, S. S. (1999). Analysis of a large structure/biological activity data set using recursive partitioning. *Journal of Chemical Information and Computer Sciences*, **39**, 1017–1026.

SAS System (2003), Version 8.2. SAS Institute, North Carolina.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 486–494.

Shi, P. and Tsai, C.-L. (2002). Regression model selection—A residual likelihood approach. *Journal of the Royal Statistical Society B*, **64**, 237–252.

Sittampalam, G. S., Iversen, P. W., Boadt, J. A., Kahl, S. D., Bright, S., Zock, J. M., Janzen, W. P., and Lister, M. D. (1997). Design of signal windows in high throughput screening assays for drug discovery. *Journal of Biomolecular Screening*, **2**, 159–169.

Tukey, J. W. (1994). Reminder sheets for "Allowances for various types of error rate". In *The Collected Works of John W. Tukey, volume VIII, Multiple Comparisons: 1948–1983.*. Editor: H. I. Braun, pages 335–339, Chapman & Hall, New York.

Tukey, J. W. (1997). More honest foundations for data analysis. *Journal of Statistical Planning and Inference*, **57**: 21–28.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience, New York.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*, second edition. Springer Verlag, New York.

Warmuth, M. K., Liao, J., Ratsch, G., Mathieson, M., Putta, S., and Lemmenk, C. (2003). Active learning with support vector machines in the drug discovery process. *Journal of Chemical Information and Computer Sciences*, **43**, 667–673.

Weston, J., Perez-Cruz, F., Bousquet, O., Chapelle, O., Elisseeff, A., and Schölkopf, B. (2002). Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics*, **1**, 1–8.

Wikel, J. H. and Higgs, R. E. (1997). Point: Applications of molecular diversity analysis in high throughput screening. *Journal of Biomolecular Screening*, **2**, 65–66.

World Drug Index (2002). Thompson Derwent, London.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120–131.

Young, S. S., Ekins, S., and Lambert, C. G. (2002). So many targets, so many compounds, but so few resources. *Current Drug Discovery*, 1–6 (www.currentdrugdiscovery.com).

Zemroch, P. J. (1986). Cluster analysis as an experimental design generator. *Technometrics*, **28**, 39–49.

# 5
# Design and Analysis of Screening Experiments with Microarrays

Paola Sebastiani, Joanna Jeneralczuk, and Marco F. Ramoni

Microarrays are an important exploratory tool in many screening experiments. There are multiple objectives for these experiments including the identification of genes that change expression under two or more biological conditions, the discovery of new cellular or molecular functions of genes, and the definition of a molecular profile that characterizes different biological conditions underlying, for example, normal or tumor cells. The technology of microarrays is first described, followed by some simple comparative experiments and some of the statistical techniques that are used for their analysis. A very important question arising in the design of screening experiments with microarrays is the choice of the sample size and we describe two approaches to sample size determination. The first approach is based on the concept of reproducibility, and the second uses a Bayesian decision-theoretic criterion to make a trade-off between information gain and experiment costs. Finally some of the open problems in the design and analysis of microarray experiments are discussed.

## 1 Introduction

One of the results of the Human Genome project is that we now know that the human DNA comprises between 30,000 and 35,000 genes. Only about 50% of these genes have known functions and several projects around the world are currently underway to characterize these newly discovered genes and to understand their role in cellular processes or in mechanisms leading to disease.

An avenue of research focuses on *gene expression*, that is, the process by which a gene transcribes the genetic code stored in the DNA into molecules of mRNA that are used for producing proteins. The measurement of the expression levels of all the genes in a cell is nowadays made possible by the technology of microarrays (Lockhart and Winzeler, 2000). The basic idea underlying the technology of microarrays is that the genes responsible for different biological conditions may have different expression and hence produce molecules of mRNA in differing proportions. Microarray technology allows the measurement of the expression levels of all the genes in a cell, thus producing its *molecular profile*. By measuring the molecular profiles of cells in different conditions, researchers can identify the

genes responsible for the different biological conditions as those with different expression levels, or *differential expression*.

An important use of microarray technology is the generation of scientific hypotheses: many microarray experiments are conducted in order to discover new genes that may have a role in a particular biological process or may be responsible for disease. Because of their high costs, however, microarray experiments are often limited in sample size. From the experimental design point of view, the use of microarray technology as a tool for generating hypotheses raises novel design and methodology issues. Even the design of a simple experiment conducted to discover the molecular profiles of two biological conditions opens up basic issues such as the choice of the minimum sample size required to make a reliable claim about a hypothesis.

In Sections 2 to 4, we review the technology of synthetic oligonucleotide microarrays and describe some of the popular statistical methods that are used to discover genes with differential expression in simple comparative experiments. A novel Bayesian procedure is introduced in Section 5 to analyze differential expression that addresses some of the limitations of current procedures. We proceed, in Section 6, by discussing the issue of sample size and describe two approaches to sample size determination in screening experiments with microarrays. The first approach is based on the concept of reproducibility, and the second approach uses a Bayesian decision-theoretic criterion to trade off information gain and experimental costs. We conclude, in Section 7, with a discussion of some of the open problems in the design and analysis of microarray experiments that need further research.

## 2 Synthetic Oligonucleotide Microarrays

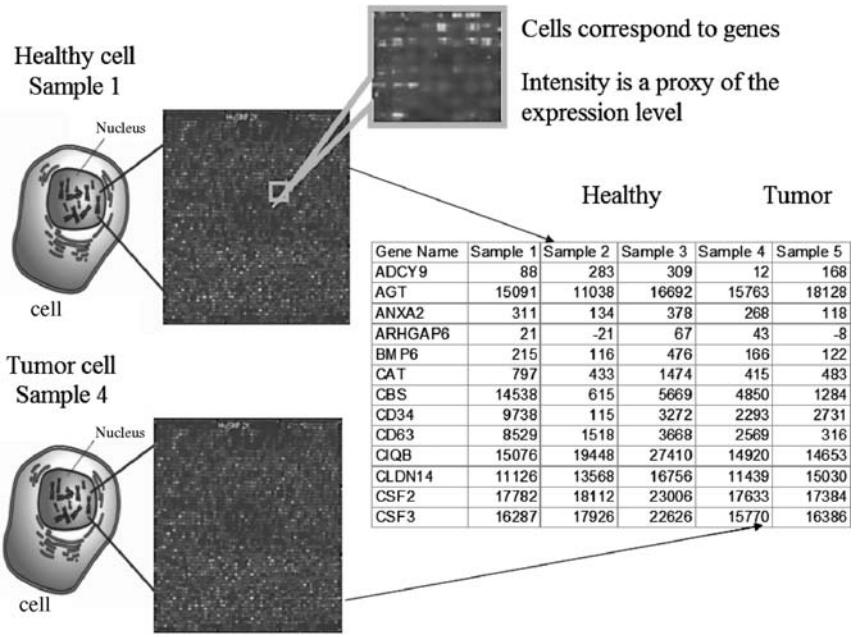
The modern concept of gene expression dates back to the seminal work of Jacob and Monod (1961) and their fundamental discovery that differential gene expression determines different protein abundance that induces different cell functions. During its expression, a gene transcribes its DNA sequence combining the nucleotides $A$, $T$, $C$, and $G$ into molecules of mRNA (messenger ribonucleic acid) and these are then transported out of the cell nucleus and used as a template for making a protein. This two-step representation of the protein-synthesis process constitutes the *central dogma of molecular biology* (Crick, 1970).

Because the first step of a gene expression consists of copying its DNA sequence into mRNA molecules, the resulting proportion of mRNA molecules provides a quantitative measure of the gene expression level. Thus, the expression level of all genes in a cell can be measured by the mRNA abundance of each gene. This is achieved by exploiting a property of the DNA sequence and the mRNA molecule produced during the gene expression: each pair of molecules binds together at a particular temperature. This property is known as *hybridization* (Lennon and Lehrach, 1991).

There are different technologies for microarrays and we refer the reader to Chapter 6 and the review given by Sebastiani et al. (2003) for a description of cDNA microarrays. Here, we focus on *synthetic oligonucleotide microarrays*. Technically, a synthetic oligonucleotide microarray is a gridded platform where each location of the grid corresponds to a gene and contains several copies of a short specific DNA segment that is characteristic of the gene (Duggan et al., 1999). The short specific segments are known as *synthetic oligonucleotides* and the copies of synthetic oligonucleotides that are fixed on the platform are called the *probes*.

The rationale behind synthetic oligonucleotide microarrays is based on the concept of probe redundancy; that is, a set of well-chosen probes is sufficient to identify a gene uniquely. Therefore, synthetic oligonucleotide microarrays represent each gene by a set of probes unique to the DNA of the gene. On the GeneChip® platform, each probe consists of a segment of DNA, and each gene is represented by a number of *probe pairs* ranging from 11 in the Human Genome U133 set to 16 in the Murine Genome U74v2 set and the Human Genome U95v2. A probe pair consists of a perfect match probe and a mismatch probe. Each perfect match probe is chosen on the basis of uniqueness criteria and proprietary empirical rules designed to improve the odds that probes will hybridize to mRNA molecules with high specificity. (Specificity, here, means the hybridization of the mRNA molecules in the target to the probes corresponding to the correct genes, and only to those.) The mismatch probe is identical to the corresponding perfect match probe except for the nucleotide in the central position, which is replaced with its complementary nucleotide, so $A$ is replaced by $T$ and vice versa, and $C$ is replaced by $G$ and vice versa. The inversion of the central nucleotide makes the mismatch probe a further specificity control because, by design, hybridization of the mismatch probe can be attributed to either nonspecific hybridization or background signal caused by the hybridization of salts to the probes (Lockhart et al., 1996). Each cell-grid of an Affymetrix oligonucleotide microarray consists of millions of samples of a perfect match or mismatch probe, and the probes are scattered across the microarray in a random order to avoid systematic bias.

To measure the expression level of the genes in a cell, investigators prepare the *target* by extracting the mRNA from the cell and making a fluorescencetagged copy. This tagged copy is then hybridized to the probes in the microarray. During the hybridization, if a gene is expressed in the target cells, its mRNA representation will bind to the probes on the microarray, and its fluorescence tagging will make the corresponding probe brighter. Studies have demonstrated that the brightness of a probe has a high correlation with the amount of mRNA in the original sample. Therefore, the measure of each probe intensity is taken as a proxy of the mRNA abundance for the corresponding gene in the sample, and a robust average (such as Tukey's biweights average; see Hoaglin et al., 2000, Chapter 11) of the intensities of the probe set determines a relative expression for the corresponding gene. Full details are in the Affymetrix document describing the statistical algorithm that is available from www.affymetrix.com/support/technical/whitepapers and a summary is given by Sebastiani et al. (2003). Figure 1 shows the three steps of a microarray experiment.




| Gene Name | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| ADCY9 | 88 | 283 | 309 | 12 | 168 |
| AGT | 15091 | 11038 | 16692 | 15763 | 18128 |
| ANXA2 | 311 | 134 | 378 | 268 | 118 |
| ARHGAP6 | 21 | -21 | 67 | 43 | -8 |
| BMP6 | 215 | 116 | 476 | 166 | 122 |
| CAT | 797 | 433 | 1474 | 415 | 483 |
| CBS | 14538 | 615 | 5669 | 4850 | 1284 |
| CD34 | 9738 | 115 | 3272 | 2293 | 2731 |
| CD63 | 8529 | 1518 | 3668 | 2569 | 316 |
| CIQB | 15076 | 19448 | 27410 | 14920 | 14653 |
| CLDN14 | 11126 | 13568 | 16756 | 11439 | 15030 |
| CSF2 | 17782 | 18112 | 23006 | 17633 | 17384 |
| CSF3 | 16287 | 17926 | 22626 | 15770 | 16386 |

FIGURE 1. A diagram of a microarray experiment. The mRNA in a cell is fluorescently labeled and hybridized to the microarray. After the hybridization, the intensity of each probe is captured into an image that is then processed to produce a proxy of the expression level of each gene in the target. In this figure, five microarrays were used to measure the molecular profiles of three healthy cells (Samples 1–3) and two tumor cells (Samples 4 and 5). (Image courtesy of Affymetrix.)

## 3 Design of Comparative Experiments

A typical microarray experiment produces the expression levels of thousands of genes in two or more biological conditions (described below). We denote the measured expression levels by $y = \{y_{kji}\}$, where the index $k$ specifies the $k$th gene in the microarray ($k = 1, \ldots, g$) and the index $i$ denotes the $i$th sample measured in condition $j$. Because of technical and biological variability that is due to difficulties in the execution of the experiment and variability between different tissues used to extract the mRNA, more than one sample in each biological condition is usually measured. We denote by $n_j$ the number of samples measured in condition $j$ so that $i = 1, \ldots, n_j$ ($j = 1, \ldots, m$). Note that samples of the same biological condition may be pure replications or biological replications. In the first situation, the target hybridized to the microarrays is made of mRNA extracted from the same cell and, in the second case, the target hybridized to the microarrays is made of mRNA extracted from different cells.

We call the set of expression levels measured for a gene across different conditions its *expression profile* and we use the term *sample molecular profile* (or

*sample*) to denote the expression level of the genes measured with one microarray in a particular condition. Formally, the expression profile of a gene $k$ in condition $j$ is the set of measurements $y_{kj} = \{y_{kj1}, \ldots, y_{kjn_j}\}$, the overall expression profile of the same gene across all conditions is the set $y_k = \{y_{k1}, y_{k2}, \ldots, y_{km}\}$ and the $i$th sample profile of condition $j$ is the set of measurements $y_{ji} = \{y_{1ji}, \ldots, y_{gji}\}$.

Common experimental objectives are the identification of the genes with significant differential expression in two or more conditions, and the development of models that can classify new samples on the basis of their molecular profiles. In some experiments, the conditions may be controllable experimental factors such as doses of a drug or the time point at which to conduct the experiment. In general observational studies, which account for a large proportion of microarray studies, the experimenter defines the conditions of interest (often disease and normal tissues) and measures the molecular profile of samples that are randomly selected. The study designs are typically *case-control* (Schildkraut, 1998) with subjects selected according to their disease status: that is, "cases" are subjects affected by the particular disease of interest, whereas "controls" are unaffected by the disease. For example, in an experiment conducted to identify the genes that are differentially expressed between normal lung cells and tumor lung cells, tissues from unaffected and affected patients are randomly chosen and each tissue provides the mRNA sample that is hybridized to the microarray.

In observational studies, the main design issue is the choice of the sample size, whereas sample size determination and treatment choice are the primary design issues in factorial experiments. Sample size determination depends on the analytical method used to identify the genes with different expression and the optimality requirements selected for the study. These topics are examined in the next two sections.

## 4 Analysis of Comparative Experiments

Popular techniques for identifying the genes with different expression in two biological conditions labeled 1 and 2 are based on the $t$-statistic:

$$t_k = \frac{\bar{y}_{k1} - \bar{y}_{k2}}{SE(\bar{y}_{k1} - \bar{y}_{k2})} \quad \text{for } k = 1, \ldots, g, \tag{1}$$

where $\bar{y}_{kj}$ is the mean expression level of gene $k$ in condition $j$, and the standard error of the sample mean difference, $SE(\bar{y}_{k1} - \bar{y}_{k2})$, is computed assuming that there will be different variances for the two conditions. Because of the large variability of gene expression data measured with microarrays, several authors have suggested some forms of penalization for the denominator of the $t$-statistic. For example, Golub et al. (1999) suggested that the standard error $SE(\bar{y}_{k1} - \bar{y}_{k2})$ should be replaced by the quantity

$$s_{S2Nk} = \frac{s_{k1}}{\sqrt{n_1}} + \frac{s_{k2}}{\sqrt{n_2}},$$

where $s_{kj}$ is the sample standard deviation of the expression levels of gene $k$ in condition $j$. The ratio

$$|\bar{y}_{k1} - \bar{y}_{k2}|/s_{S2Nk} \tag{2}$$

is known as the *signal-to-noise ratio*. Other forms of penalization are justified by the fact that the standard error may be very small for genes with small expression values, thus inflating the value of the $t$-statistic. Based on this reasoning, Tusher et al. (2000) suggested adjustment of the standard error to $a + SE(\bar{y}_{k1} - \bar{y}_{k2})$ where the constant $a$ is chosen to minimize the coefficient of variation of the $t$-statistic of all the genes. More recently, Efron et al. (2001) suggested replacement of $a$ by the 90th percentile of the standard errors $SE(\bar{y}_{k1} - \bar{y}_{k2}),\ k = 1, \ldots, g$.

The choice of the threshold for selecting the genes with a statistically significant change of expression is often based on distribution-free methods. The main idea is to compute the value of a statistic from the data in which the sample labels that represent the conditions are randomly reshuffled. By repeating this process a large number of times, it is possible to construct the empirical distribution of a statistic under the null hypothesis of no differential expression. From this distribution one can select a gene-specific threshold for rejecting the null hypothesis with a particular significance level. Authors have also developed algorithms for multiple comparison adjusted $p$-values (see, for example, Dudoit et al., 2001; also, see Chapter 6).

Distribution-free methods tend to be widely used in practice, but they often require a large sample size to detect the genes with different expression, while achieving a small false positive rate (Zien et al., 2003). Some authors have suggested making distribution assumptions on the gene expression data and the most popular choice is to assume that gene expression data follow a lognormal distribution (Baldi and Long, 2001; Ibrahim et al., 2002). Another stream of work focuses on the estimation of the *fold change* of expression, that is, the ratio of the sample means assuming a Gamma distribution for the gene expression data (Chen et al., 1997; Newton et al., 2001). We investigated the adequacy of these distributional assumptions on some large data sets available from http://www-genome.wi.mit.edu/cancer and none of these distributions appear to be, by themselves, appropriate for all genes.

For example, Figure 2 depicts the histogram of one sample of size 50 of the probe set corresponding to the "HSYUBG1 Homo sapiens ubiquitin" gene in the U95Av2 Affymetrix microarray. The distribution in Figure 2(a) has an exponential decay, with a long right tail. Figure 2(b) displays the distribution of the log-transformed data and shows that, under the log-transformation, the left skewness of the original data is removed by introducing a right skewness. This phenomenon is typically observed when the log-transformation is applied to data that follow a Gamma distribution and, consequently, bias is introduced in the estimation of the mean (McCullagh and Nelder, 1989).

The probe set of Figure 2 was selected from a publicly available data set of expression profiles comprising 50 normal prostatectomy samples and 52 tumor prostatectomy samples (Singh et al., 2002). We tested the competing distribution

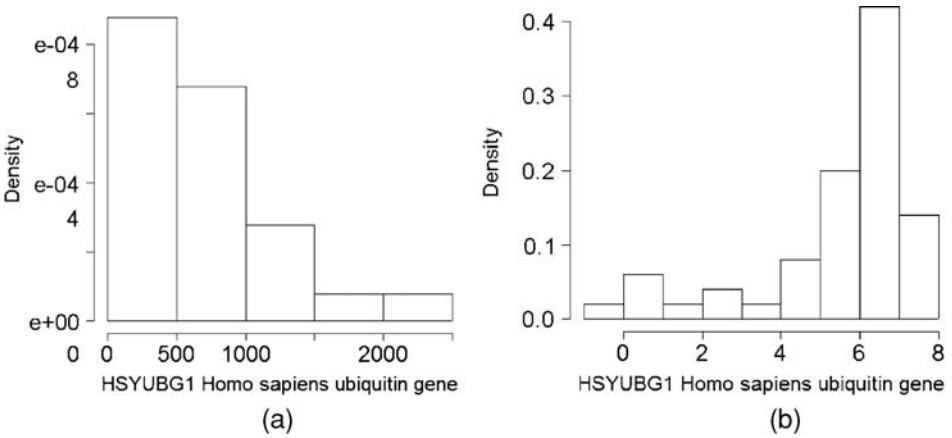



FIGURE 2. Distribution of gene expression data from 50 prostatectomy samples measured with the U95Av2 Affymetrix microarray: (a) histogram of the expression level of the "HSYUBG1 Homo sapiens ubiquitin gene"; (b) histogram of the same gene expression level after the log-transformation was used.

assumptions on each of the 12,625 probe sets using the likelihood ratio test described by Jackson (1969), at a 5% significance level. About 50% of gene expression data sets appeared to be better described by lognormal distributions, whereas the remaining 50% were better described by Gamma distributions. This finding opens a serious issue because discriminating between lognormal and Gamma distributions is notoriously difficult, particularly in small samples (Jackson, 1969). To overcome this issue, we developed a methodology for differential analysis that uses model averaging to account for model uncertainty.

## 5 Bayesian Analysis of Differential Gene Expression

Our software for Bayesian analysis of differential gene expression (BADGE), uses model averaging to solve the problem of model uncertainty in gene expression data. BADGE measures the differential expression by the fold change $\theta_k$. Formally, if we let $\mu_{kj}$ denote the average expression level for the gene $k$ in condition $j (j = 1, 2)$, the fold change is the ratio

$$\theta_k = \frac{\mu_{k1}}{\mu_{k2}}, \quad k = 1, \ldots, g, \tag{3}$$

where $g$ is the number of genes. No change of expression for gene $k$ is represented by $\theta_k = 1$, whereas changes of expression are represented by a fold change $\theta_k < 1$ and $\theta_k > 1$.

The method implemented in BADGE is Bayesian. It regards the fold change $\theta_k$ as a random variable so that the differential expression of each gene is measured by the posterior probability, $p(\theta_k > 1|y_k)$, given that the observed differential expression

profile was $y_k$ (defined in Section 3). Values of $p(\theta_k > 1|y_k)$ near 0.5 identify the genes that do not change expression across the two conditions, whereas values of $p(\theta_k > 1|y_k)$ near 1 identify the genes that have larger expression in condition 1 than in condition 2, and values of $p(\theta_k > 1|y_k)$ near 0 identify the genes that have smaller expression in condition 1 than in condition 2. The posterior probability of differential expression of a gene $k$ is independent of the measurements of the other genes because we assume that the expression values of different genes are independent given the parameter values. This assumption may not be realistic because genes are known to interact with each other, but it allows screening for genes with differential expression. More advanced methods that take gene–gene dependence into account are described by Sebastiani et al. (2004).

The software, BADGE, computes the posterior probability of differential expression of each gene by assuming Gamma or lognormal distributions (with certain probabilities) for each of the gene expression data sets. It then applies the technique known as *Bayesian model averaging*, described by, for example, Hoeting et al. (1999), as follows.

If we let $M_{Lk}$ and $M_{Gk}$ denote the model assumptions that the expression data of gene $k$ follow either a lognormal or a Gamma distribution, the posterior probability $p(\theta_k > 1|y_k)$ can be computed as

$$p(\theta_k > 1|y_k) = p(\theta_k > 1|M_{Lk}, y_k)p(M_{Lk}|y_k) + p(\theta_k > 1|M_{Gk}, y_k)p(M_{Gk}|y_k), \tag{4}$$

where $p(\theta_k > 1|M_{Lk}, y_k)$ and $p(\theta_k > 1|M_{Gk}, y_k)$ are the posterior probabilities of differential expression assuming a lognormal and a Gamma model, respectively. The weights $p(M_{Lk}|y_k)$ and $p(M_{Gk}|y_k) = 1 - p(M_{Lk}|y_k)$ are the posterior probabilities of the two models. Because a Bayesian point estimate of the fold change is the expected value $E(\theta_k|y_k)$ of the posterior distribution of $\theta_k$, the point estimate of the fold-change $\theta_k$ is computed by averaging the point estimates conditional on the two models

$$E(\theta_k|y_k) = E(\theta_k|M_{Lk}, y_k)p(M_{Lk}|y_k) + E(\theta_k|M_{Gk}, y_k)p(M_{Gk}|y_k). \tag{5}$$

Similarly, an approximate $(1 - \alpha)\%$ credible interval (see, for example, Berger, 1985, Chapter 1) is computed by averaging the credible limits computed under the two models. In particular, if $(l_{kL}, u_{kL})$ and $(l_{kG}, u_{kG})$ are the $(1 - \alpha)\%$ credible limits conditional on the two models, an approximate $(1 - \alpha)\%$ credible interval for $\theta_k$ is $(\theta_{kl}, \theta_{ku})$, where

$$\theta_{kl} = l_{kL}\,p(M_{Lk}|y_k) + l_{kG}\,p(M_{Gk}|y_k),$$
$$\theta_{ku} = u_{kL}\,p(M_{Lk}|y_k) + u_{kG}\,p(M_{Gk}|y_k).$$

Details of these calculations are given in Appendix A. To select the subset of genes that characterizes the molecular profile of two experimental conditions, we proceed as follows. The posterior probability of differential expression $p(\theta_k > 1|y_k)$ is the probability that the gene $k$ has larger expression in condition 1 than in condition 2, given the available data. If we fix a threshold $s$ so that we select

as differentially expressed those genes that have $p(\theta_k > 1|y_k) < s$ and $p(\theta_k < 1|y_k) < s$, then the expected number of genes selected by chance would be $2(g \times s)$, where $g$ is the number of genes in the microarray. If this number is fixed to be $c$, then the threshold $s$ is $c/(2g)$. This can be interpreted as the expected error rate, that is, the proportion of genes with differential expression that fail to be detected.

# 6 Sample Size Determination

A crucial question in the design of comparative experiments is the determination of the sample size that is sufficient for drawing conclusions from the data with some level of confidence. The traditional approach to sample size determination is power-based and leads to the choice of the sample size needed to achieve a desired power for rejecting the null hypothesis in favor of a particular alternative hypothesis. Dow (2003) and Zien et al. (2003) have investigated this approach in simulation studies and showed that the sample size depends on the minimum fold change to be detected, the statistical method used for the estimation of the fold change, and the trade-off between false-positive and false-negative rates. So, for example, with two conditions, a minimum of 25 samples per condition is needed in order to detect genes that have more than a twofold change with a false-positive rate of 0.1% and a power of 80% using the standard $t$-test (see Zien et al., 2003). However, this approach appears to be too restrictive for a screening experiment and it is also strongly dependent upon debatable assumptions about the distribution of gene expression data. Therefore, we introduce two different criteria based on the concept of reproducibility and information gain.

## *6.1 Reproducibility*

The first approach to sample size determination that we investigate is based on the concept of *reproducibility*. The objective is to identify the minimum sample size that is needed to reproduce, with high probability, the same results in other similar experiments. To investigate this issue, we need a large database of microarray experiments from which we can select nonoverlapping subsets of data that are analyzed with the same statistic. The reproducibility is then measured by computing the agreement between the statistic values for the different subsets. A measure of agreement is the scaled correlation $(1 + \rho_i)/2$, where $\rho$, is the average correlation between statistics in $q_i$ samples of size $n_i$. Suppose, for example, the differential expression of a gene $k$ in two biological conditions is measured by the $t$-statistics $t_k(D_{ji})$ defined in (1), where $D_{ji}$ is the data set of size $n_i$ used in the comparison. As we repeat the analysis in nonoverlapping data sets of the same size $n_i$, we obtain the set of values $t(D_{ji}) = \{t_1(D_{ji}), \ldots, t_g(D_{ji})\}$ for $j = 1, \ldots, q_i$, and we can measure the pairwise agreement by the $q_i(q_i - 1)/2$ correlations

$$\rho_{rs,i} = corr(t(D_{ri}), t(D_{si})).$$

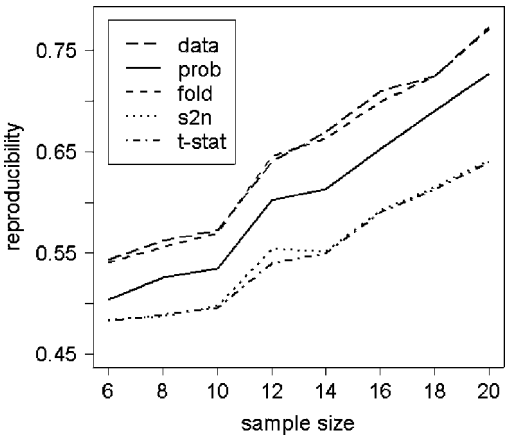



FIGURE 3. Reproducibility of: the posterior probability (solid line); the estimate of the fold change (dashed line); the $t$-statistic (dash-dot line); the signal-to-noise ratio (dotted line); together with the data reproducibility between sample means (long dashed line), for different sized samples. Reproducibility is $(1 + \rho_i)/2$, where $\rho_i$ is the average correlation between statistics in samples of size $n_i (= 6, \ldots, 20)$.

The average correlation $\rho_i$ is then computed by averaging the $q_i(q_i - 1)/2$ pairwise correlations.

Similar calculations can be done using other statistics. For example, Figure 3 shows a plot of the reproducibility of the posterior probability (4) and the estimate $\theta_k$ of the fold change (3) computed by BADGE together with the reproducibility of the $t$-statistic (1) and of the signal-to-noise ratio statistic (2) implemented in the GeneCluster software. The long-dashed line reports the *data reproducibility* that was measured by the scaled correlation between the ratios of sample means. To measure the reproducibility, we selected 32 nonoverlapping subsets from the large data set of 102 expression profiles of prostatectomy samples described in Section 4. Specifically, we chose eight different sample sizes ($n_i = 6, 8, 10, 12, 14, 16, 18, 20$) and, for each of the eight sample sizes $n_i$, we created four data sets by selecting four sets of $n_i$ normal samples and $n_i$ tumor samples from the original database. This procedure generated 32 data sets. Then we used BADGE to compute the posterior probability of differential expression and the estimate of the fold change $\hat{\theta}_k$ in each data set. We also analyzed the data sets with GeneCluster using the standard $t$ and signal-to-noise ratio statistics.

The plot in Figure 3 shows a substantially larger reproducibility of the fold change and posterior probability computed by BADGE compared to the $t$ and signal-to-noise ratio statistics. Furthermore, the reproducibility of the estimated fold change is virtually indistinguishable from the data reproducibility. However, the reproducibility of the posterior probability is about 5% less than the reproducibility of the data, and both the $t$ and signal-to-noise ratio statistics are about 10% less reproducible than the data. The data reproducibility of experiments with fewer than 10 samples per group is very low (below 60%). A reproducibility higher than 70% requires at least 20 samples per group. To investigate further the effect of sample size on the reproducibility of detecting differential expression, we examined the reproducibility of the analysis with 1329 genes that were selected by BADGE with

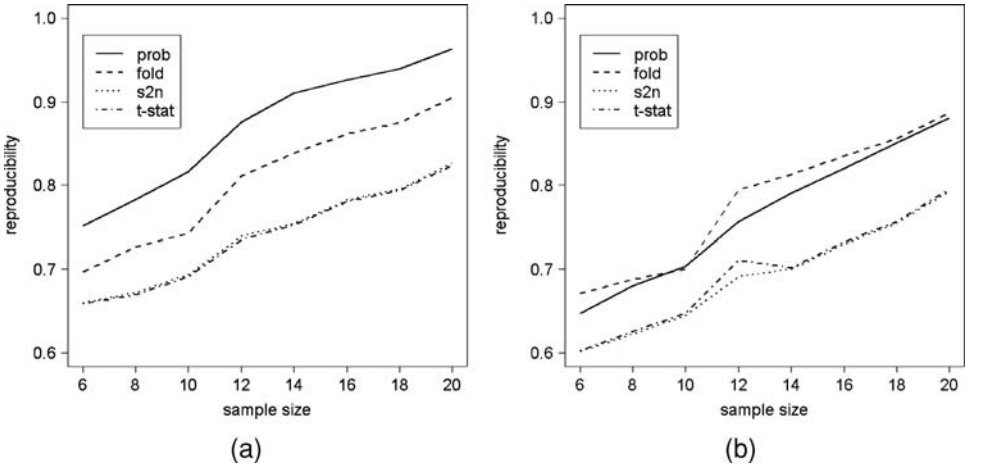



FIGURE 4. (a) Reproducibility of: the posterior probability (solid line); the estimate of the fold change (dashed line); the $t$-statistic (dash-dot line) and the signal-to-noise ratio (dotted line) for different sample sizes, for the 1329 genes selected as most differentially expressed by BADGE on the whole data set; (b) the same analysis for the 1329 genes selected as most differentially expressed by the $t$-statistic. The reproducibility is measured by $(1 + \rho_i)/2$, where $\rho_i$ is the average correlation between statistics in samples of size $n_i (= 6, \ldots, 20)$.

posterior probability of $\theta_k > 1$ being smaller than 0.01 or larger than 0.99 in the whole data set comprising 102 samples.

The objective of this comparison was to investigate whether these genes would be detected as differentially expressed in experiments with smaller sample sizes. Figure 4(a) summarizes the results and we can see the large reproducibility of the analysis for small sample sizes: the reproducibility is above 70% even in experiments with only 6 samples per group, and above 80% when the number of samples per group is at least 12. Once again, the reproducibility of the fold analysis conducted by BADGE is consistently larger than that of the analysis conducted with the $t$ or signal-to-noise ratio statistics. We also repeated the analysis using about 1300 genes that were selected by values of the $t$-statistic smaller than $-2$ or larger than 2 in the whole data set. The results are summarized in Figure 4(b) and show that the selection of the gene by the $t$-statistic is 5% less reproducible compared to the selection based on BADGE. These results suggest the need for at least 12 samples per condition to have substantial reproducibility with BADGE, whereas the analysis based on the $t$ or signal-to-noise ratio statistics would require more than 20 samples per condition.

## 6.2 *Average Entropy*

Sample size determination based on reproducibility does not take the experimental costs into account. In this section, we introduce a formal decision-theoretic

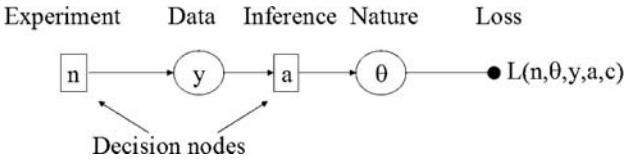



FIGURE 5. A decision tree for the choice of sample size and inference method. The first decision node represents the choice of the sample size $n$. After this decision, the experiment is conducted and generates the data $y$ that are assumed to follow a distribution with parameter $\theta$. The data are used to make an inference on the parameter $\theta$, and the second decision node $a$ represents the statistical procedure that is used to make this inference. The last node represents the loss induced by choosing such an experiment.

approach that allows us to choose the sample size by trading off between the gain of information provided by the experiment with the experimental costs.

The decision problem is represented by the decision tree in Figure 5, in which open circles represent chance nodes, squares represent decision nodes, and the black circle is a value node. The first decision node is the selection of the sample size $n$ used in the experiment, and $c$ represents the cost per observation. The experiment will generate random data values $y$ that have to be analyzed by an inference method $a$. The difference between the true state of nature, represented by the fold changes $\theta = (\theta_1, \ldots, \theta_g)$, and the inference will determine a loss $L(\cdot)$ that is a function of the two decisions $n$ and $a$, the data, and the experimental costs. There are two choices in this decision problem: the optimal sample size and the optimal inference.

The solutions are found by "averaging out" and "folding back" (Raiffa and Schlaifer, 1961), so that we compute the expected loss at the chance nodes (open circles), given everything to the left of the node. We determine the best actions by minimizing the expected loss at the decision nodes. The first decision is the choice of the inference method $a$ and the optimal decision $a^*$ (or Bayes action) is found by minimizing the expected loss $E\{L(n, \theta, y, a, c)\}$, where the expectation is with respect to the conditional distribution of $\theta$ given $n$ and $y$. The expected loss evaluated in the Bayes action $a^*$ is called the *Bayes risk* and we denote it by

$$R(n, y, a^*, c) = E\{L(e, \theta, y, a^*, c)\}.$$

This quantity is also a function of the data $y$, and the optimal sample size is chosen by minimizing the expected Bayes risk $E\{R(n, y, a^*, c)\}$, where this expectation is with respect to the marginal distribution of the data.

A popular choice for the loss function $L(\cdot)$ is the log-score defined as

$$L(n, \theta, y, a, c) = -\log a(\theta|n, y) + nc, \tag{6}$$

in which $a(\theta|y, n)$ is a distribution for the parameter $\theta$, given the data and the sample size $n$. This loss function was originally advocated by Good (1952) as a proper measure of uncertainty conveyed by a probability distribution. Lindley

(1956, 1997) proposed the use of this loss function to measure the information gain provided by an experiment and to determine the optimal sample size for an experiment. With this choice of loss function, the Bayes action $a^*$ is the posterior distribution $p(\theta|n, y)$ of $\theta$ given $n$ and $y$, and so the Bayes risk is given by

$$\begin{aligned} R(n, y, a^*, c) &= -\int [\log p(\theta|n, y)] p(\theta|n, y) d\theta + nc \\ &\equiv Ent(\theta|n, y) + nc. \end{aligned} \tag{7}$$

The quantity

$$Ent(\theta|n, y) = -\int [\log p(\theta|n, y)] p(\theta|n, y) d\theta \tag{8}$$

is known as the Shannon entropy, or entropy, and the negative Shannon entropy represents the amount of information about $\theta$ contained in the posterior distribution. Therefore, the negative Bayes risk in (7) represents the trade-off between information gain and experimental costs.

As stated above, in order to choose the optimal sample size $n = n_1 + n_2$, we need to minimize with respect to $n_1$ and $n_2$ the expected Bayes risk (where $n_1$ and $n_2$ are the numbers of tissue samples for conditions 1 and 2); that is,

$$E\{R(n, y, a^*, c)\} = \int Ent(\theta|n, y) p(y|n) dy + nc.$$

Because we assume that expression data are independent given the parameters, the joint posterior density of the parameter vector $\theta$ is

$$p(\theta|n, y) = \prod_k p(\theta_k|n, y_k).$$

This independence implies that the entropy $Ent(\theta|n, y)$ is the sum of the entropies $\sum_k Ent(\theta_k|n, y_k)$, and so the expected Bayes risk becomes

$$E\{R(n, y, a^*, c)\} = \sum_k \int Ent(\theta_k|n, y_k) p(y_k|n) dy_k + nc.$$

In the software BADGE, we account for model uncertainty by averaging the results of the posterior inference conditional on the Gamma and lognormal distributions for the gene expression data. As a parallel with the sample size determination when the inference process is based on model averaging, we therefore introduce the *Average Entropy*, denoted by $Ent_a(\cdot)$, and defined as

$$\begin{aligned} Ent_a(\theta_k|n, y_k) = {} & p(M_{Lk}|n, y_k) Ent(\theta_k|n, y_k, M_{Lk}) \\ & + p(M_{Gk}|n, y_k) Ent(\theta_k|n, y_k, M_{Gk}). \end{aligned}$$

This quantity averages the Shannon entropies conditional on the Gamma and lognormal models, with weights given by their posterior probabilities. In Appendix B, we show that the average entropy is a concave function on the space of probability distributions which is monotone under contractive maps (Sebastiani and

Wynn, 2000) and has some nice decomposition properties. These properties ensure that

$$Ent_a(\theta|n, y) = \sum_k Ent_a(\theta_k|n, y_k).$$

This last expression allows us to simplify the calculation of the expected Bayes risk $E\{R(n, y, a^*, c)\}$ by describing it as an average of Bayes risks conditional on the Gamma and lognormal models, with weights given by their prior probabilities:

$$\begin{aligned} E\{R(n, y, a^*, c)\} &= \sum_k E\{Ent_a(\theta_k|n, y_k)\} + nc \\ &= \sum_k p(M_{Lk}) \int p(y_k|n, M_{Lk}) Ent(\theta_k|n, y_k, M_{Lk}) dy_k \\ &\quad + \sum_k p(M_{Gk}) \int p(y_k|n, M_{Gk}) Ent(\theta_k|n, y_k, M_{Gk}) dy_k + nc. \end{aligned}$$

The importance of this result is that it leads to an overall objective criterion for sample size determination that averages criteria based on specific model assumptions. Thus it provides a solution that is robust to model uncertainty. Closed-form calculations of (8) are intractable, so we have developed numerical approximations to the conditional entropies $Ent(\theta_k|n, y_k, M_{Lk})$ and $Ent(\theta_k|n, y_k, M_{Gk})$. The computations of the expected Bayes risk are performed via stochastic simulations and the exact objective function is estimated by curve fitting as suggested by Müller and Parmigiani (1995). These details are available on request from the authors.

Figure 6 shows an example of the stochastic estimation of the Bayes risk as a function of the sample sizes $n_1$ and $n_2$, where the data were obtained by resampling from the data set of 102 prostatectomy samples described in Section 6.1. From the results on the reproducibility, we estimated that a sample of size $n$ induces a reproducibility of about $(22.5 \log(n) - 4)\%$, so we used as loss function

$$-\log(p(\theta_k|n, y_k)) + .22 * \log(n) - .04.$$

An interesting fact is the sharp decrease of the estimated Bayes risk when the sample size increases from six to ten samples per condition, whereas the reduction in risk is less effective for larger sample sizes (see Figure 6). This result agrees with the findings in Section 6.1 about the reproducibility of the analysis. Furthermore, the effect of changing the number of samples in the two conditions is not symmetric. This finding is intriguing and suggests that, at least in microarray experiments which compare normal versus tumor samples, it is better to have a larger number of normal samples than tumor samples. An intuitive explanation of this finding is that tumor samples are less variable because the individuals are all affected by the same disease.

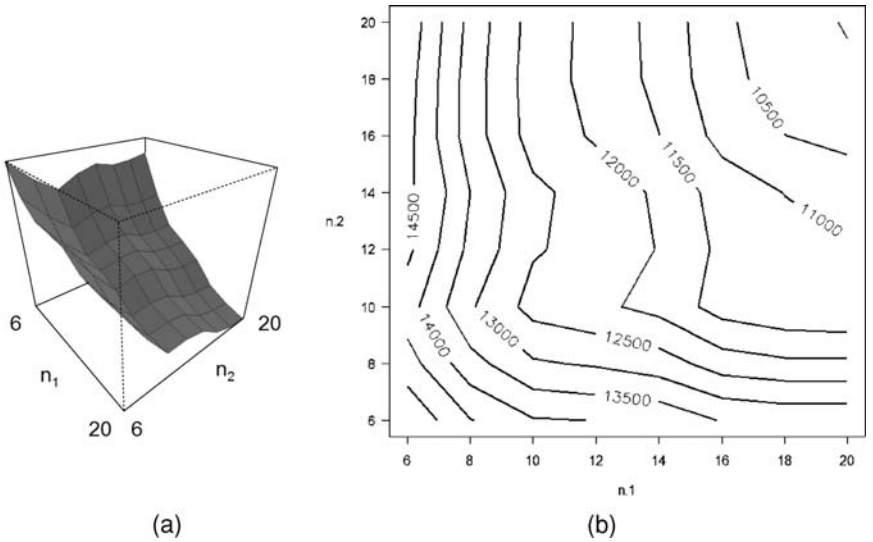



FIGURE 6. (a) The surface for the estimated Bayes risk (vertical axis) as a function of the number of samples $n_1$ and $n_2$ for the two conditions; (b) the contour plot of the same surface.

## 7 Discussion

This chapter has focused on the design of comparative experiments conducted to identify genes with differential expression. However, microarrays are used for broader experimental objectives and their use challenges statisticians with novel design questions. In comparative experiments, an important question is whether it is better to make pure replicates of the expression measurements of the same cell or to collect more data from different cells (biological replicates). Lee et al. (2000) showed that a single replicate is not sufficient to achieve reproducible results and suggested that at least three pure replications of each measurement should be used. The costs of microarray experiments still impose serious sample size limitations, and the designer of the experiment needs to trade off the number of biological replications against the number of pure replications. The best solution depends on the objective of the analysis. If the objective is to have an accurate estimate of the *error variability* in the microarray measurements, then an experiment with a large number of pure replicates and a small number of biological replicates will be preferable to an experiment with one observation of each biological sample. However, in experiments in which the *biological variability* between samples is expected to be large, such as in clinical studies involving human subjects, the investment of resources in biological replicates rather than in pure replicates is, intuitively, the best strategy. This dilemma in the design of an experiment and the

lack of an "out of-the-box" answer shows the need for further research in this area.

Sample size and treatment choice are key design questions for general multifactor experiments. Authors have proposed the use of standard factorial experiments in completely randomized designs, block designs, or Latin squares (see, for example, Chapter 6 and Churchill, 2003). However, the unusual distribution of gene expression data makes one question the relevance of standard orthogonal factorial experiments in this context.

Another important problem that has received little attention in the design community is the development of design criteria for experiments which are not limited to the estimation of particular parameters. For example, data from comparative experiments are often used to define classification models that are able to predict a clinical feature by using the molecular profile of cells in a tissue. This objective is particularly important for cancer classification (see Golub et al., 1999) when is is difficult to discriminate between different subtypes of cancer. The typical approach is to select the genes with differential expression and use them to build a classification model. Several models have been proposed in the literature and an overview is given by Sebastiani et al. (2003). Validation of the classification accuracy is carried out by using a training set of data to build the model and a test set of data to assess the classification accuracy of the model. Here, an important design question is the sample size needed to determine a classification model that is sufficiently accurate; an interesting approach based on learning curves is described in Mukherjee et al. (2003).

More complex are the design issues involved in microarray experiments conducted to identify gene functions or their network of interactions. The assumption that genes with similar functions have similar expression patterns underlies the popular approach of clustering gene expression profiles and sample molecular profiles for identifying subgroups of genes with similar expression patterns in a subset of the samples (see, for example, Eisen et al., 1998). Design issues are not only the sample size determination but also the selection of the time points at which to make the measurements in temporal experiments. When the goal of the experiment is to model the network of gene interactions, we move into the area of experimental design for causal inference. Popular formalisms of knowledge representation, such as Bayesian networks (Cowell et al., 1999) and dynamic Bayesian networks, seem to be the ideal tools for capturing the dependency structure among gene expression levels (Friedman et al., 2000; Segal et al., 2001; Yoo et al., 2002; Sebastiani et al., 2004). Apart from preliminary work by Pearl (1999) and Spirtes et al. (1999), experimental design to enable causal inference with Bayesian networks is an unexplored research area.

*Acknowledgments.* This research was supported by the NSF program in Bioengineering and Environmental Systems Division/Biotechnology under Contract ECS-0120309.

# Appendix A
# Computation of Posterior Distributions

Brief details are given below of some numerical approximations used to compute the posterior distribution of the fold change $\theta_k$ (3), for $k = 1, \ldots, g$. We assume that, given the model parameters, the expression data $y_{kji}$ are independent for different genes and samples.

## *Computation Details: Lognormal Distribution*

Suppose the expression data $y_{kji}$ are generated from a random variable $Y_{kj}$ that follows a lognormal distribution with parameters $\eta_{kj}$ and $\sigma^2_{kj}$ so that the mean $\mu_{kj}$ is $e^{\eta kj+\sigma^2_{kj}/2}$ and the variance is $\mu^2_{kj}(e^{\sigma^2_{kj}} - 1)$. In particular, $X_{kj} = \log(Y_{kj})$ is normally distributed with mean $\eta_{kj}$ and variance $\sigma^2_{kj}$. Because

$$\begin{aligned} p(\theta_k > 1|M_{Lk}, y_k) &= p(\log(\mu_{k1}) - \log(\mu_{k2}) > 0|M_{Lk}, y_k) \\ &= p(\eta_{k1} - \eta_{k2} + (\sigma^2_{k1} - \sigma^2_{k2})/2 > 0|M_{Lk}, y_k) \end{aligned}$$

any inferences about $\theta_k$ can be done equivalently on the parameters $\eta_{kj}, \sigma^2_{kj}$ of the log-transformed variables. The posterior probability $p(\theta_k > 1|M_{Lk}, y_k)$ can be computed as

$$\begin{aligned} p(\theta_k > 1|M_{Lk}, y_k) = \int p(\eta_{k1} - \eta_{k2} > \frac{\sigma^2_{k2} - \sigma^2_{k1}}{2}|\sigma^2_{k1}, \sigma^2_{k2}, M_{Lk}, y_k) \\ \times f(\sigma^2_{k1}, \sigma^2_{k2}|M_{Lk}, y_k)d\sigma^2_{k1}d\sigma^2_{k2}, \end{aligned} \tag{9}$$

where $f(\sigma^2_{k1}, \sigma^2_{k2}|M_{Lk}, y_k)$ denotes the posterior density of the parameters $\sigma^2_{k1}, \sigma^2_{k2}$.

We assume a standard uniform prior on $\eta_{kj}$ and $\log(\sigma^2_{kj})$ and prior independence of $(\eta_{k1}, \sigma^2_{k1})$ from $(\eta_{k2}, \sigma^2_{k2})$. Then, it is well known that, given the data, the parameters $\sigma^2_{k2}, \sigma^2_{k1}$ are independent and distributed as $s^2_{kj}/\sigma^2_{kj} \sim \chi^2_{n_i-1}$, $i = 1, 2$, where $\chi^2_n$ denotes a $\chi^2$ distribution on $n$ degrees of freedom, and

$$s^2_{kj} = \sum_j (x_{kji} - \bar{x}_{kj})^2/(n_i - 1)$$

is the sample variance of the log-transformed data $x_{kji} = \log(y_{kji})$ in condition $i$. Similarly, $\eta_{kj}|\sigma^2_{kj}$ is normally distributed with mean $\bar{x}_{kj}$ and variance $\sigma^2_{kj}/n_i$, and the marginal distribution of $\eta_{kj}$ is

$$\eta_{kj} \sim (s^2_i/n_i)^{1/2} t_{n_i-1} + \bar{x}_{kj},$$

where $t_n$ is a Student's $t$ distribution on $n$ degrees of freedom (Box and Tiao, 1973).

To compute the integral in (9), we use the fact that, for fixed $\sigma^2_{k2}, \sigma^2_{k1}$, the quantity $p(\eta_{k1} - \eta_{k2} > (\sigma^2_{k2} - \sigma^2_{k1})/2)$ is the cumulative distribution function of a standard normal distribution evaluated at

$$z = -\{(\sigma^2_{k2} - \sigma^2_{k1})/2 - (\bar{x}_{k1} - \bar{x}_{k2})\}/\sqrt{\sigma^2_{k1}/n_1 + \sigma^2_{k2}/n_2}$$

and then we average this quantity with respect to the joint posterior distribution of $\sigma_{k2}^2, \sigma_{k1}^2$. Because there does not seem to be a closed form solution, we use a two-step numerical approximation. First we approximate the integral in (9) by the first-order approximation

$$p(\eta_{k1} - \eta_{k2} > \frac{s_{k2}^2 - s_{k1}^2}{2} | M_{Lk}, y_k)$$

and then we use the numerical approximation to the Behrens–Fisher distribution described by Box and Tiao (1973), to approximate the posterior probability by

$$p(\theta_k > 1 | M_{Lk}, y_k) \approx p\left(t_b > -\frac{\bar{x}_{k1} - \bar{x}_{k2} + s_{k1}^2/2 - s_{k2}^2/2}{a(s_{k1}^2/n_1 + s_{k2}^2/n_2)^{1/2}}\right).$$

The scaling factor $a$ and the adjusted degrees of freedom $b$ are given in Box and Tiao (1973). For large $n_1, n_2$, the scaling factor $a$ approaches 1 and the degrees of freedom $b$ approach $n_1 + n_2 - 2$ so that the posterior distribution of $\eta_{k1} - \eta_{k2}$ is approximately the noncentral Student's $t$-statistic

$$(s_{k1}^2/n_1 + s_{k2}^2/n_2)^{1/2} t_{n_1+n_2-2} + \bar{x}_{k1} - \bar{x}_{k2}.$$

The approximation is applicable for $n_1$ and $n_2$ greater than 5, and the comparisons we have conducted against inference based on MCMC methods have shown that this approximation works well for samples of size 6 or more.

An approximate estimate of the fold change $\theta_k$ is

$$\hat{\theta}_k = \exp(\bar{x}_{k1} - \bar{x}_{k2} + s_{k1}^2/2 - s_{k2}^2/2)$$

and approximate credible limits are given by

$$l_{kL} = \exp((\bar{x}_{k1} - \bar{x}_{k2} + s_{k1}^2/2 - s_{k2}^2/2) - t_{(1-\alpha/2,b)} a(s_{k1}^2/n_1 + s_{k2}^2/n_2)^{1/2}),$$
$$u_{kL} = \exp((\bar{x}_{k1} - \bar{x}_{k2} + s_{k1}^2/2 - s_{k2}^2/2) - t_{(1-\alpha/2,b)} a(s_{k1}^2/n_1 + s_{k2}^2/n_2)^{1/2}),$$

where $t_{(1-\alpha/2,b)}$ is the $(1 - \alpha/2)$ quantile of a Student's $t$ distribution on $b$ degrees of freedom.

## *Computation Details: Gamma Distribution*

Suppose now that the gene expression data follow a Gamma distribution with parameters $\alpha_{kj}, \beta_{kj}$ that specify the mean and the variance of the distribution as $\mu_{kj} = \alpha_{kj}/\beta_{kj}$ and $V(Y_{kj}|\alpha_{kj}, \beta_{kj}) = \mu_{kj}^2/\alpha_{kj}$. We wish to compute the posterior distribution of $\theta_k = \mu_{k1}/\mu_{k2}$, or equivalently

$$\theta_k = \frac{\alpha_{k1}}{\alpha_{k2}} \frac{\beta_{k2}}{\beta_{k1}}.$$

If $\alpha_{kj}$ were known, say $\alpha_{kj} = \hat{\alpha}_{kj}$, then the use of a uniform prior for $\beta_{kj}$ would determine the posterior distribution for $\beta_{kj}|y_k$, namely, Gamma$(n_i\hat{\alpha}_{kj} + 1, n_i\bar{y}_{kj})$. The value $\hat{\alpha}_{kj}$ can be, for example, the maximum likelihood estimate of $\alpha_{kj}$, which

is the solution of the equation:

$$f(\alpha_{kj}) = \log(\alpha_{kj}) - \psi(\alpha_{kj}) - \log(\bar{y}_{kj}) + \sum_j \log(y_{kji})/n_i = 0,$$

where $\psi(\alpha) = d\log(\Gamma(\alpha))/d\alpha$ is the digamma function. Then it is easily shown that $2n_i\bar{y}_{kj}\beta_{kj} \sim \chi^2_{2(n_i\hat{\alpha}_{kj}+1)}$ (Casella and Berger, 1990). Furthermore, $\beta_{k1}|y_k$ and $\beta_{k2}|y_k$ are independent and, because the ratio of two independent random variables that have $\chi^2$ distributions is proportional to an F distribution (Box and Tiao, 1973), the distribution of the ratio $\beta_{k2}/\beta_{k1}$ is easily found to be

$$\frac{\beta_{k2}}{\beta_{k1}} \sim \left(\frac{n_1\bar{y}_{k1}}{n_2\bar{y}_{k2}}\right)\left(\frac{n_2\hat{\alpha}_{k2}+1}{n_1\hat{\alpha}_{k1}+1}\right) F_{2(n_2\hat{\alpha}_{k2}+1),2(n_1\hat{\alpha}_{k1}+1)}$$

and an approximation to the probability $p(\theta_k > 1|M_{Gk}, y_k)$ is

$$p(\theta_k > 1|M_{Gk}, y_k) = p\left(F_{2(n_2\hat{\alpha}_{k2}+1),2(n_1\hat{\alpha}_{k1}+1)} > \frac{\bar{y}_{k2}}{\bar{y}_{k1}} \cdot \frac{\hat{\alpha}_{k2}}{\hat{\alpha}_{k1}} \cdot \frac{\hat{\alpha}_{k1}+1/n_1}{\hat{\alpha}_{k2}+1/n_2}\right).$$

The point estimate for $\theta_k$ is given by $\hat{\theta}_k = \bar{y}_{k1}/\bar{y}_{k2}$, and $(1-\alpha)\%$ credible limits are

$$l_{kG} = \frac{\bar{y}_{k1}}{\bar{y}_{k2}} \cdot \frac{\hat{\alpha}_{k1}}{\hat{\alpha}_{k2}} \cdot \frac{\hat{\alpha}_{k2}+1/n_2}{\hat{\alpha}_{k1}+1/n_1} f_{\alpha/2,2(n_2\hat{\alpha}_{k2}+1),2(n_1\hat{\alpha}_{k1}+1)},$$

$$u_{kG} = \frac{\bar{y}_{k1}}{\bar{y}_{k2}} \cdot \frac{\hat{\alpha}_{k1}}{\hat{\alpha}_{k2}} \cdot \frac{\hat{\alpha}_{k2}+1/n_2}{\hat{\alpha}_{k1}+1/n_1} f_{1-\alpha/2,2(n_2\hat{\alpha}_{k2}+1),2(n_1\hat{\alpha}_{k1}+1)}.$$

The assessment of the error of the approximation depends on the posterior variance of $\alpha_{kj}$ for which we do not have a closed form expression. Empirical comparisons that we conducted on gene expression data sets suggest that the results based on our numerical approximation are virtually indistinguishable from those obtained by Markov chain Monte Carlo methods when $n_1, n_2 > 10$. Details are described by Sebastiani et al. (2005).

## *Computation Details: Mixing Weights*

To compute the mixing weights in equations (4) and (5) in Section 5, we assume that changes in the average expression levels between the two conditions can at most affect the parameter values but not the distribution membership. Therefore, the mixing weights are the posterior probabilities $p(M_{Lk}|y_k)$ and $p(M_{Gk}|y_k)$ computed by disregarding the distinction between the two conditions $j = 1, 2$. We use the approximation to the posterior odds $B_k = p(M_{Lk}|y_k)/p(M_{Gk}|y_k)$ given by the Bayesian information criterion to make the choice independent of the prior probabilities (Kass and Raftery, 1995). In this way, the posterior probability $p(M_{Lk}|y_k)$ is $B_k/(1+B_k)$ and $p(M_{Gk}|y_k) = 1/(1+B_k)$. The Bayesian information criterion is essentially the likelihood ratio:

$$B_k = \frac{p(M_{Lk}|y_k)}{p(M_{Gk}|y_k)} = \frac{f_l(y_k|\hat{\eta}_k, \hat{\sigma}^2_k)}{f_g(y_k|\hat{\alpha}_k, \hat{\beta}_k)}, \tag{10}$$

where $f_L(y_k|\hat{\eta}_k, \hat{\sigma}_k^2)$ and $f_G(y_k|\hat{\alpha}_k, \hat{\beta}_k)$ are the likelihood functions for the lognormal and Gamma models evaluated in the maximum likelihood estimates $\hat{\eta}_k, \hat{\sigma}_k^2, \hat{\alpha}_k, \hat{\beta}_k$ of the parameters. See Sebastiani et al. (2005) for further details.

# Appendix B
# Properties of the Average Entropy

In this appendix, we prove some general properties of the average entropy in the context of gene expression analysis. We denote by $\theta$ the change of expression of a generic gene in two conditions, and we suppose that the expression values follow either a Gamma distribution, $M_G$, or a lognormal distribution, $M_L$. In this case, the average entropy becomes:

$$Ent_a(\theta) = w_1 Ent(\theta|M_L) + (1 - w_1)Ent(\theta|M_G),$$

where, for simplicity of notation, $w_1$ denotes the probability of the model $M_L$, and $1 - \omega_1$ is the probability of the model $M_G$. The quantities $Ent(\theta|M_L)$ and $Ent(\theta|M_G)$ denote, respectively, the Shannon entropy of $\theta$ computed under the assumption that the gene expression data follow a lognormal and a Gamma distribution.

**Theorem 1 (Concavity).** *The average entropy $Ent_a(\theta)$ is a concave function of the set of probability distributions for $\theta$.*

Proof. The result follows from the fact that Shannon entropy is concave in the space of probability distribution (DeGroot, 1970), and the average entropy is a convex combination of Shannon entropies.

**Theorem 2 (Monotonicity).** *Let $\eta = \psi(\theta)$ be a smooth transformation of $\theta$, such, that $\psi^{-1}$ exists, and let $J$ be the Jacobian of the transformation $\psi^{-1}$. Then*

$$\begin{cases} Ent_a(\eta) > Ent_a(\theta), \; if |J| < 1; \\ Ent_a(\eta) < Ent_a(\theta), \; if |J| > 1. \end{cases}$$

Proof. The result follows from the monotonicity of Shannon entropy (Sebastiani and Wynn, 2000).

**Theorem 3 (Decomposability).** *The average entropy of the random vector $\theta = \{\theta_1, \theta_2\}$ can be decomposed as*

$$Ent_a(\theta_1, \theta_2) = Ent_a(\theta_1) + E_{\theta_1}\{Ent_a(\theta_2|\theta_1)\}.$$

Proof. Let $M_{L1}$ and $M_{L2}$ denote lognormal distributions for the expression values of two genes, and let $w_1$ and $w_2$ be the posterior probability assigned to the models $M_{L1}$ and $M_{L2}$. When we decompose the average entropy of $\theta_1$ and $\theta_2$ we need to consider the space of model combinations

$$\mathcal{M} = \{(M_{L1}, M_{L2}), (M_{L1}, M_{G2}), (M_{G1}, M_{L2}), (M_{G1}, M_{G2})\}.$$

If we assume that the model specifications are unrelated, and that expression values of different genes are independent given the parameter values, then the probability distribution over the model space $\mathcal{M}$ is $w_1w_2$, $w_1(1-w_2)$, $(1-w_1)w_2$, $(1-w_1)(1-w_2)$. Then we have

$$\begin{aligned} Ent_a(\theta_1, \theta_2|\mathcal{M}) = {} & w_1w_2Ent(\theta_1, \theta_2|M_{L1}, M_{L2}) \\ & + w_1(1-w_2)Ent(\theta_1, \theta_2|M_{L1}, M_{G2}) \\ & + (1-w_1)w_2Ent(\theta_1, \theta_2|M_{G1}, M_{L2}) \\ & + (1-w_1)(1-w_2)Ent(\theta_1, \theta_2|M_{G1}, M_{G2}). \end{aligned}$$

By the property of Shannon entropy $Ent(\theta_1, \theta_2) = Ent(\theta_1) + E_{\theta_1}\{Ent(\theta_2|\theta_1)\}$, where $E_\theta(\cdot)$ denotes expectation with respect to the distribution of $\theta$, it follows that

$$\begin{aligned} & w_1w_2Ent(\theta_1, \theta_2|M_{L1}, M_{L2}) \\ & \quad = w_1w_2Ent(\theta_1, |M_{L1}) + w_1w_2E_{\theta_1|M_{L1}}\{Ent(\theta_2|\theta_1, M_{L2})\} \end{aligned}$$

and, similarly,

$$\begin{aligned} & w_1(1-w_2)Ent(\theta_1, \theta_2|M_{L1}, M_{G2}) \\ & \quad = w_1(1-w_2)Ent(\theta_1|M_{L1}) + w_1(1-w_2)E_{\theta_1|M_{L1}}\{Ent(\theta_2|\theta_1, M_{G2})\}; \\ & (1-w_1)w_2Ent(\theta_1, \theta_2|M_{G1}, M_{L2}) \\ & \quad = (1-w_1)w_2Ent(\theta_1|M_{G1}) + (1-w_1)w_2E_{\theta_1|M_{G1}}\{Ent(\theta_2|\theta_1, M_{L2})\}; \\ & (1-w_1)(1-w_2)Ent(\theta_1, \theta_2|M_{G1}, M_{G2}) = (1-w_1)(1-w_2) \\ & \quad \times Ent(\theta_1|M_{G1} + (1-w_1)(1-w_2)E_{\theta_1|M_{G1}}\{Ent(\theta_2|\theta_1, M_{G2})\}. \end{aligned}$$

Now group the terms

$$w_1w_2Ent(\theta_1|M_{L1}) + w_1(1-w_2)Ent(\theta_1|M_{L1}) = w_1Ent(\theta_1|M_{L1})$$

and

$$\begin{aligned} & (1-w_1)w_2Ent(\theta_1|M_{G1}) + (1-w_1)(1-w_2)Ent(\theta_1|M_{G1}) \\ & \quad = (1-w_1)Ent(\theta_1|M_{G1}) \end{aligned}$$

to derive

$$w_1Ent(\theta_1|M_{L1}) + (1-w_1)Ent(\theta_1|M_{G1}) = Ent_a(\theta_1).$$

Similarly, we can group the terms

$$\begin{aligned} & w_1E_{\theta_1|M_{L1}}\{w_2Ent(\theta_2|\theta_1, M_{L2}) + (1-w_2)Ent(\theta_2|\theta_1, M_{G2})\} \\ & \quad = w_1E_{\theta_1|M_{L1}}\{Ent_a(\theta_2|\theta_1)\} \end{aligned}$$

and

$$\begin{aligned} & (1-w_1)E_{\theta_1|M_{G1}}\{w_2Ent(\theta_2|\theta_1, M_{L2}) + (1-w_2)Ent(\theta_2|\theta_1, M_{G2})\} \\ & \quad = (1-w_1)E_{\theta_1|M_{G1}}\{Ent_a(\theta_2|\theta_1)\}, \end{aligned}$$

to derive

$$w_1 E_{\theta_1|M_{L1}}\{Ent_a(\theta_2|\theta_1)\} + (1-w_1)E_{\theta_1|M_{G1}}\{Ent_a(\theta_2|\theta_1)\} = E_{\theta_1}\{Ent_a(\theta_2|\theta_1)\}$$

that concludes the proof.

**Theorem 4 (Additivity).** *If $\theta_1, \theta_2$ are independent, then*

$$Ent_a(\theta_1, \theta_2) = Ent_a(\theta_1) + Ent_a(\theta_2).$$

Proof. The result follows from the previous theorem.

## *References*


Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, second edition. Springer, New York.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. John Wiley and Sons, New York.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Duxbury Press, Belmont.

Chen, Y., Dougherty, E., and Bittner, M. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, **2**, 364–374.

Churchill, G. (2003). Comment to "Statistical challenges in functional genomics". *Statistical Science*, **18**, 64–69.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

Crick, F. H. C. (1970). Central dogma of molecular biology. *Nature*, **227**, 561–563.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

Dow, G. S. (2003). Effect of sample size and *p*-value filtering techniques on the detection of transcriptional changes induced in rat neuroblastoma (NG108) cells by mefloquine. *Malaria journal*, **2**.
Available from *www.malariajournal.com/content/2/1/4*.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2001). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics*, **21**, 10–14.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151–1160.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the U.S.A.*, **95**, 14863–14868.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S.

(1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society B*, **14**, 107–114.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (2000). *Understanding Robust and Exploratory Data Analysis*. John Wiley and Sons, New York.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, **14**, 382–417.

Ibrahim, J. G., Chen, M. H., and Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, **97**, 88–99.

Jackson, O. A. Y. (1969). Fitting a gamma or log-normal distribution to fibre-diameter measurements on wool tops. *Applied Statistics*, **18**, 70–75.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, **3**, 318–356.

Kass, R. E. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Lee, M. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the U.S.A.*, **18**, 9834–9839.

Lennon, G. G. and Lehrach, H. (1991). Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics*, **7**, 314–317.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, **27**, 986–1005.

Lindley, D. V. (1997). The choice of sample size. *Journal of the Royal Statistical Society D*, **46**, 129–138.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. (1996). Expression monitoring by hybridization to high-density oligonu-cleotide arrays. *Nature Biotechnology*, **14**, 1675–1680.

Lockhart, D. and Winzeler, E. (2000). Genomics, gene expression, and DNA arrays. *Nature*, **405**, 827–836.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. Chapman & Hall, London.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology*, **10**, 119–142.

Müller, P. and Parmigiani, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association*, **90**, 1322–1330.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, **8**, 37–52.

Pearl, J. (1999). Graphs, structural models, and causality. In *Computation, Causation, and Discovery*, 95–140. MIT Press, Menlo Park.

Raiffa, H. A. and Schlaifer, R. S. (1961). *Applied Statistical Decision Theory*. MIT Press, Cambridge.

Schildkraut, J. M. (1998). Examining complex genetic interactions. In *Gene Mapping in Complex Human Diseases*, 379–410. John Wiley and Sons, New York.

Sebastiani, P., Abad, M., and Ramoni, M. F. (2005). Bayesian networks for genomic analysis. In *Genomic Signal Processing and Statistics*, Hindawi Publishing Corporation. 281–320.

Sebastiani, P., Gussoni, E., Kohane, I. S., and Ramoni, M. F. (2003). Statistical challenges in functional genomics (with discussion). *Statistical Science*, **18**, 33–70.

Sebastiani, P., Xie, H., and Ramoni, M. F. (2005). *Bayesian analysis of comparative microarray experiments by model averaging*. Submitted (available upon request)

Sebastiani, P. and Wynn, H. P. (2000). Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society B*, **62**, 145–157.

Segal, E., Taskar, B., Gasch, A., Friedman, N., and Koller, D. (2001). Rich probabilistic models for gene expression. *Bioinformatics*, **1**, 1–9.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., DAmico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Spirtes, P., Glymour, C., Scheines, R., Meek, C., Fienberg, S., and Slate, E. (1999). Prediction and experimental design with graphical causal models. In *Computation , Causation, and Discovery*, 65–94. MIT Press, Menlo Park.

Tusher, V. G., Tibshirani, R., and Chu, G. (2000). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the U.S.A.*, **98**, 5116–5121.

Yoo, C., Thorsson, V., and Cooper, G. (2002). Discovery of causal relationships in a gene-regulation pathway form a mixture of experimental and observational DNA microarray data. In *Proceedings of the Pacific Symposium on Biocomputing*. Available from *http://psb.stanford.edu*.

Zien, A., Fluck, J., Zimmer, R., and Lengauer, T. (2003). Microarrays: How many do you need? *Journal of Computational Biology*, **10**, 653–667.

# 6
# Screening for Differential Gene Expressions from Microarray Data

Jason C. Hsu, Jane Y. Chang, and Tao Wang

Living organisms need proteins to provide structure, such as skin and bone, and to provide function to the organism through, for example, hormones and enzymes. Genes are *translated* to proteins after first being *transcribed* to messenger RNA. Even though every cell of an organism contains the full set of genes for that organism, only a small set of the genes is functional in each cell. The levels at which the different genes are functional in various cell types (their *expression levels*) can all be screened simultaneously using microarrays. The design of two-channel microarray experiments is discussed and ideas are illustrated through the analysis of data from a designed microarray experiment on gene expression using liver and muscle tissue. The number of genes screened in a microarray experiment can be in the thousands or tens of thousands. So it is important to adjust for the multiplicity of comparisons of gene expression levels because, otherwise, the more genes that are screened, the more likely incorrect statistical inferences are to occur. Different purposes of gene expression experiments may call for different control of multiple comparison error rates. We illustrate how control of the statistical error rate translates into control of the rate of incorrect biological decisions. We discuss the pros and cons of two forms of multiple comparisons inference: testing for significant difference and providing confidence bounds. Two multiple testing principles are described: closed testing and partitioning. Stepdown testing, a popular form of gene expression analysis, is shown to be a shortcut to closed and partitioning testing. We give a set of conditions for such a shortcut to be valid.

## 1 Introduction

### *1.1 Uses of Gene Expression Analysis*

Living organisms need proteins to provide structure, such as skin and bone, and to provide function to the organism through, for example, hormones and enzymes. Genes are *translated* to proteins after first being *transcribed* to mRNA (messenger RNA). Even though every cell of an organism contains the full set of genes for that organism, only a small set of the genes is functional in each cell. The level to which the different genes are functional in various cell types (their *expression levels*) can all be screened simultaneously using microarrays.

Microarray experiments may be observational studies whose sole purpose is to screen the genome for differential gene expressions or they may have more specific aims such as the following.

- *Designer medicine*: one potential use of gene expression is to tailor medicine or treatment to individuals, with gene expressions used as explanatory or predictor variables. Microarrays may be used to provide genetic profiles for individual patients, to classify patients into more refined disease categories, or to predict an individual's predisposition to a certain disease.
- *Patient targeting*: the populations from which the samples of, for example, blood or tissue, are drawn might be two subpopulations of patients, those who have and those who have not had adverse reactions to a particular molecular entity. Currently, some molecular entities of potential benefit to many patients cannot be approved as drugs because a small but significant percentage of the patients experience adverse events (AE) when receiving the drug. If the subpopulation that is prone to adverse reactions can be identified in terms of differential gene expressions then, by not giving the drug to this subpopulation, such a molecular entity can perhaps be approved in order to benefit the vast majority of patients.
- *Drug discovery*: the populations from which the samples are drawn may be healthy and disease tissues. A possible use of the comparison of gene expressions is to find protein targets that might intervene with the disease process.

We believe the statistical formulation of gene expression analysis should reflect the intended use of microarray data. For example, to fabricate a diagnostic or prognostic microarray for designer medicine, genes might first be screened for differential expressions. Diagnostic/prognostic chips for gene profiling might then be built containing the genes that appear to be differentially expressed. Clustering (unsupervised learning) algorithms may be applied to group together genes that have similarities. Classification (supervised learning) algorithms may be applied to place patients into appropriate categories. When too many genes are present, the processes of classification and clustering may be overwhelmed, or at least may prove counterproductive. So one approach towards designer medicine is to use multiple testing first to screen for relevant genes. The analysis of gene expression data from the initial screening microarray study should thus reflect the need for the fabricated chip to meet desired requirements for *sensitivity*, probability of correctly inferring presence of a disease, and *specificity*, the probability of not inferring presence of a disease when the disease is absent; see, also, Chapter 5, Section 5.2.

On the other hand, an aim of drug discovery is to find proteins that can intervene with a disease process. For example, it might be possible to determine and synthesize the protein made by a gene that is under-expressed in disease tissues; that is, the gene is transcribing less protein than it would for normal tissue. Genes themselves are turned on and off by gene regulatory proteins (called *transcription factors*). So it might also be possible to find proteins to regulate genes that are too active, or not active enough, under disease conditions. Although a detailed

description of the transcription and translation process is given by Alberts et al. (2002, Chapter 6), for example, a good introduction is the online *DNA from the Beginning* tutorial available at http://www.dnaftb.org/dnaftb. The function of a gene can be studied by mutating or deleting it in entire organisms and then observing the consequences, but this is a slow and costly process: knocking out one gene in one mouse might cost a thousand dollars. Therefore, the statistical design of microarray experiments and the analysis of gene expression data should maximize the chance of biologists subsequently discovering the proteins that can affect a disease process.

In this chapter, our discussion is in terms of multiple testing for drug discovery. We use the example of screening genes in order to find proteins (transcription factors) that might cure a disease or alleviate its symptoms to illustrate how to couple the gene expression analysis with the intended purpose of the microarray experiment.

## *1.2 Types of Microarrays*

There are two main types of microarrays: oligonucleotide microarrays and two-channel microarrays. An overview of the concepts behind the two types of microarrays can be found in Module 36 of the *DNA from the Beginning* tutorial. Two-channel microarrays were invented by Pat Brown in the mid 1990s. Institutions and researchers can purchase robotic machines in order to design and make their own two-channel microarrays.

The DNA sequence of a gene contains both coding segments (called exons) and noncoding segments (called introns) and only the coding segments are transcribed into mRNA. In making a two-channel microarray, the first step is to prepare target DNA that contain only the coding segments of known and purported genes. These segments can consist of either entire sequences or shorter segments called expressed sequence tags (EST). These target DNA are then deposited (spotted) onto glass slides in a grid pattern. From each of two different mRNA samples (for example, normal and disease tissues), the complementary DNA (cDNA) that can base pair with the target DNA is then prepared. This process is called *reverse transcription*. One sample is dyed fluorescent red and the other sample is dyed fluorescent green. The samples are then allowed to bind (*hybridize*) to the target DNA of the known genes and the ESTs spotted on the microarray. A scanner subsequently reads separately the red and green intensity of each spot on the microarray and produces, for the gene spotted there, a measure of its expression in each of the two samples.

In contrast to two-channel microarrays, oligonucleotide arrays are single-channel microarrays, meaning that each microarray provides a profile of the gene expressions of a single sample. Institutions and researchers purchase ready-made oligonucleotide microarrays from manufacturers who build them using proprietary technologies. For a detailed description of oligonucleotide microarrays, see Chapter 5.

An advantage of being able to design one's own microarrays is that traditional genomics and new technology can work more effectively together. For example,

one can use linkage analysis or quantitative trait loci (QTL) analysis to narrow the search for genes involved in a disease to a relatively small number of genes, and then spot only these genes on two-channel microarrays. This allows for more replications of the relevant genes on each array while keeping the total number of spots on each array the same and this, in turn, leads to sharper statistical inferences on each gene of relevance.

### *1.3 Design of Microarray Experiments*

Variations observed in gene expressions may be attributable to differences in the conditions being compared in the samples (for example, normal and disease tissues), or to differences in nuisance factors such as array, dye, and spot. Thus, we believe the design of microarrays should follow the statistical principles of blocking, randomization, and replication.

Blocking for known nuisance factors removes systematic bias in estimation. For example, it has been observed that different dyes may have different labeling efficiencies. The use of a *dye-swap* design, which reverses the assignment of dyes to the conditions being compared on successive arrays, avoids confounding the difference due to efficiency of dye labeling with the difference in the conditions being compared. If microarrays are used for diagnostic or prognostic purposes, unbiased estimation is an important safeguard for the public.

Blocking can also improve the precision of estimation if within-block variation is small compared to between-block variation. Sample-to-sample variability and variation in the amounts deposited on different arrays can be blocked effectively by hybridizing each replicate sample to a different array and considering the factor "array" as a blocking factor in modeling the data.

Randomization is the basis for valid statistical inference. Random assignment of replicate samples to arrays and dyes helps to avoid unintended systematic bias due to such nuisance factors. In designing a two-channel microarray, a software driver for the robotic arrayer allows appropriate programming to randomize and replicate the spotting of the complementary DNA (cDNA) on each array. Unfortunately, randomization of spotting on each array is often not done currently, perhaps due to inconvenience.

Replication is necessary for the estimation of variability and can be accomplished by having multiple samples and spotting each gene more than once on a microarray. Churchill (2003) gave an insightful discussion of the costs and benefits of having different kinds of replications; see Churchill (2002) and Yang and Speed (2002) for further discussions (also see Chapter 5).

## 2 Forms of Statistical Inference

There are two forms of statistical inference that are commonly used for differential gene expressions. The first is to infer the presence of differential expressions. The second is to infer the magnitude of the difference of differential expressions.

To be able to execute a statistical analysis that infers the presence of differential expressions, only a specification is needed of the probability mechanism leading to the observations under the null hypothesis of equality of expressions. One reason for the popularity of testing the equality of gene expressions is seemingly that the tests can be performed without modeling (using permutation tests, for example), that is, with no specification of dependency among the test statistics, and no specification of how differential expressions affect the responses. A minimal model is the *randomization* model (Rao, 1973, page 502) which assumes that the assignment of observed vectors of gene expressions to the two conditions has occurred at random. A test of a null hypothesis of equality of expressions can only infer the presence of differential expressions and not the absence. However, in clinical trials, it has been recognized that testing for the presence of a treatment effect is inadequate. Medical decisions are increasingly based on "clinically meaningful differences" as opposed to "statistically significant differences" (differences from zero). Examples of such inferences include establishment of the equivalence and noninferiority of treatments (see ICH E10, 1999).

To execute a statistical analysis capable of providing confidence intervals or bounds on the magnitude of differential gene expressions, the approach we take is to build a model that specifies how a disease condition might, affect the distribution of the observed gene expressions; that is, we specify a model that is more elaborate than the randomization model, connecting responses with parameters $\theta_i, i = 1, \ldots, g$, which, for example, may be differences or ratios of long run averages, or location shifts. Even if only the presence of an effect needs to be inferred, model building is still useful for constructing multiple tests that are computationally feasible (see Section 5).

A number of authors have proposed model-based approaches to the analysis of gene expression data. Churchill and Oliver (2001), Churchill (2003), Kerr and Churchill (2001a,b), and Wolfinger et al. (2001) discussed the modeling of gene expression data from controlled experiments, and Lee et. al. (2000) and Thomas et al. (2001) discussed the modeling of gene expressions data from observational studies. Our modeling approach follows that of previous authors so that we may

1. Build a model that includes blocking factors to remove nuisance effects (for example, array and dye effects);
2. Perform diagnostic checks for the appropriateness of the model;
3. Estimate differential expressions and their standard errors;

but adds the additional step:

4. Adjust for multiplicity of simultaneous inference on multiple genes based on the joint distribution of estimators.

We recommend that two types of control genes should be included in microarray experiments as described below.

Gene expressions may differ in different cell types (for example, in normal and disease tissues). These may occur even for genes not involved in the disease process (*housekeeping genes*) which are functional in all cells (see, for example,

Vandesompele et al., 2002). We recommend that housekeeping genes should be included in microarray experiments and that their observed differential expressions be used to normalize those of the genes under study. In other words, housekeeping genes can serve a purpose similar to that of *negative controls* (placebos) in clinical trials.

In microarray experiments, the observed differential gene expressions are also functions of the intensity scale that the scanner is capable of measuring. The same set of conditions will produce different amounts of differential expressions for the same gene on different scanners if the scanners have different intensity scales. With certain disease categories, there are genes known to be differentially expressed. For example, the p53 gene is mutated in most cancer tumors. We recommend that such genes be included in microarray experiments and that their observed differential expressions be used to guide the choice of what constitutes a biologically meaningful differential expression. In other words, genes known to be differentially expressed can serve a purpose similar to that of active controls in clinical trials.

## 3 Control of Error Rate

Suppose that we wish to make inferences on the parameters $\theta_i$, $i = 1, \ldots, g$, where $\theta_i$ represents the logarithm of the ratio of the expression levels of gene $i$ under normal and disease conditions. If the $i$th gene has no differential expression, then the ratio is 1 and hence $\theta_i = 0$. In testing the g hypotheses $H_{0i} : \theta_i = 0, i = 1, \ldots, g$, suppose we set $R_i = 1$ if $H_{oi}$ is rejected and $R_i = 0$ otherwise. Then, for any multiple testing procedure, one could in theory provide a complete description of the joint distribution of the indicator variables $R_1, \ldots, R_g$ as a function of $\theta_1, \ldots, \theta_g$ in the entire parameter space. This is impractical if $g > 2$. Different controls of the error rate control different aspects of this joint distribution, with the most popular being weak control of the familywise error rate (FWER), strong control of the familywise error rate, and control of the false discovery rate (FDR).

We discuss the pros and cons of controlling each error rate, in the context of drug discovery.

Weak control of the FWER, also referred to as the experimentwise error rate in some traditional statistics books, controls the maximum probability of rejecting at least one null hypothesis $H_{0i}$ when all $H_{0i}, i = 1, \ldots, g$, are true. Weak control of the FWER is inadequate in practice because, if there exists at least one differentially expressed gene, then there is no guarantee that the probability of incorrectly inferring the nondifferentially expressed genes as differentially expressed is controlled.

Strong control of the FWER controls the maximum probability of rejecting any true null hypothesis, regardless of which subset of the null hypotheses happens to be true. We call a nondifferentially expressed gene that is incorrectly inferred to be differentially expressed a *false positive*. Then controlling the FWER strongly at 5% guarantees that, with a probability of at least 95%, the number of false positives is zero, regardless of how many genes are truly differentially expressed.

Control of the FDR, controls the expected proportion of true null hypotheses rejected among all the hypotheses rejected. In the gene expressions setting, the FDR is the expected proportion of false positives among those inferred to be differentially expressed.

Whether strong control of FWER or FDR is appropriate in the analysis of microarrays may depend on how the inference on differential gene expressions is used subsequently. In one form of drug discovery, genes are first screened for differential expressions under normal and disease conditions. Subsequently, nucleotide sequences in the promoter regions of the genes selected in the first step are "mined" for unusually common sequences (called *consensus motifs*). Proteins (transcription factors) that bind to these motifs then become candidates for drug compounds to intervene with the disease process.

Although fewer genes will be selected as differentially expressed by strong control of the familywise error rate, one has confidence that each gene selected is, indeed, involved in the disease process. More genes will be selected by controlling the false discovery rate, but one is only confident that a proportion of the genes selected is involved in the disease process. In fact, the number of false positive genes selected by an FDR-controlling method is unknown, because it depends on the unknown number of genes that are truly differentially expressed. Another problem with this method is that one can manipulate the level at which the hypotheses of interest are tested by including extra null hypotheses that are already known to be false (see Finner and Roter, 2001). Suppose that a single null hypothesis is of interest. If three additional null hypotheses known to be false are tested, then the null hypothesis of interest can, in effect, be tested at a level of 20% while controlling the FDR at 5%. Thus, one can artificially inflate the chance that an FDR-controlling method will falsely discover "new" genes involved in a disease, by purposely including as many genes as possible that are already known to be involved in that disease. This is a dilemma because, as mentioned in Section 2, it is a good idea to include such genes in the study to serve as active controls on the microarrays. We thus caution against careless application of the FDR concept in gene expressions analysis.

Instead of controlling the expected proportion of incorrect rejections, one might control the expected number of incorrect rejections. As explained by Dudoit et al. (2003), in essence this is what is achieved by the significance analysis of microarrays (SAM) methods of Efron et al. (2001) and Tusher et al. (2001). The issue of which error rate is the best to control in terms of leading to more discoveries of useful transcription factors is an on-going research project.

## 4 Construction of Multiple Tests

In this section, we discuss two methods of multiple testing for the family of null hypotheses of no differential expressions

$$H_{0i} : \theta_i = 0, i = 1, \ldots, g. \tag{1}$$

Previous discussions of multiplicity adjusted testing of gene expressions, by Dudoit et al. (2002) and (2003), for example, generally took a nonmodeling approach. Because the joint distribution of the test statistics is generally not available with this approach, multiplicity adjustments in these papers tend to be calculated based on conservative inequalities (for example, the Bonferroni inequality or Sidak's inequality) or on a joint distribution of independent test statistics. In contrast, here, we describe multiplicity adjustment based on the actual joint distribution of the test statistics. However, before describing such adjustments, we first address the construction principles to which all multiple tests should adhere, regardless of the approach taken. These principles do not appear to be as well known in the field of bioinformatics as they are in clinical trials.

Tukey's method, Scheffé's method, and Dunnett's method are familiar one-step multiple comparison methods and they adjust for multiplicity based on the total number of comparisons made (see Hsu, 1996, Chapters 3 and 5). Closed testing and partition testing are two general principles that guide the construction of multiple testing methods that strongly control the FWER. In contrast to one-step testing, they are based on the idea that multiplicity needs to be adjusted only to the extent that null hypotheses may simultaneously be true, letting the data dictate the extent of multiplicity adjustment. All else being equal, closed testing and partition testing are more powerful than one-step testing. Stepdown testing, a popular form of gene expression analysis, is best thought of as a shortcut to closed/partitioning testing. We give a set of conditions in the gene expression analysis setting for such a shortcut to be valid. Subtleties of such conditions do not seem to have been fully appreciated in practice.

The closed testing principle of Marcus et al. (1976) proceeds as follows.

*Closed testing*

C1: For each subset $I \subseteq \{1, \ldots, g\}$ of genes, form the null hypotheses $H_{0I} : \{\theta_i = 0 \text{ for all } i \in I\}$.
C2: Test each $H_{0I}$ at level $\alpha$.
C3: For each $i$, infer $\theta_i \neq 0$ if and only if all $H_{0I}$ with $i \in I$ are rejected; that is, if and only if all null hypotheses containing $\theta_i = 0$ are rejected.

As an illustration, suppose that $g = 3$. Then closed testing tests the following null hypotheses.

$$
\begin{aligned}
H_{0\{1,2,3\}} &: \theta_1 = \theta_2 = \theta_3 = 0 \\
H_{0\{1,2\}} &: \theta_1 = \theta_2 = 0 \\
H_{0\{1,3\}} &: \theta_1 = \theta_3 = 0 \\
H_{0\{2,3\}} &: \theta_2 = \theta_3 = 0 \\
H_{0\{1\}} &: \theta_1 = 0 \\
H_{0\{2\}} &: \theta_2 = 0 \\
H_{0\{3\}} &: \theta_3 = 0
\end{aligned}
$$

and it is inferred, for example, that $\theta_2 \neq 0$ if and only if $H_{0\{1,2,3\}}$, $H_{0\{1,2\}}$, $H_{0\{2,3\}}$, and $H_{0\{2\}}$ are all rejected.

It turns out that a concept more easily explained and more powerful than closed testing is the partitioning principle of Stefansson et al. (1988) and Finner and Strassburger (2002), which proceeds as follows.

*Partition testing*

P1: For each subset $I \subseteq \{1, \ldots, g\}$ of genes, form the null hypotheses

$$H_{0I}^* : \{\theta_i = 0 \text{ for all } i \in I \text{ and } \theta_j \neq 0 \text{ for all } j \notin I\}.$$

P2: Test each $H_{0I}^*$ at level $\alpha$.
P3: For each $i$, infer $\theta_i \neq 0$ if and only if all $H_{0I}^*$ with $i \in I$ are rejected.

Again, suppose that $g = 3$ for illustration. Then partition testing tests:

$$\begin{aligned}
H_{0\{1,2,3\}}^* &: \theta_1 = \theta_2 = \theta_3 = 0 \\
H_{0\{1,2\}}^* &: \theta_1 = \theta_2 = 0 \text{ and } \theta_3 \neq 0 \\
H_{0\{1,3\}}^* &: \theta_1 = \theta_3 = 0 \text{ and } \theta_2 \neq 0 \\
H_{0\{2,3\}}^* &: \theta_2 = \theta_3 = 0 \text{ and } \theta_1 \neq 0 \\
H_{0\{1\}}^* &: \theta_1 = 0 \text{ and } \theta_2 \neq 0 \text{ and } \theta_3 \neq 0 \\
H_{0\{2\}}^* &: \theta_2 = 0 \text{ and } \theta_1 \neq 0 \text{ and } \theta_3 \neq 0 \\
H_{0\{3\}}^* &: \theta_3 = 0 \text{ and } \theta_1 \neq 0 \text{ and } \theta_2 \neq 0
\end{aligned}$$

and, for example, $\theta_2 \neq 0$ is inferred if and only if $H_{0\{1,2,3\}}^*$, $H_{0\{1,2\}}^*$, $H_{0\{2,3\}}^*$, and $H_{0\{2\}}^*$ are all rejected.

It is easy to see why partition testing controls the FWER:

- Because the null hypotheses $H_{0I}^*$ are disjoint, at most one $H_{0I}^*$ is true. Therefore, no multiplicity adjustment is needed in testing them to control the FWER strongly;
- Because $H_{0i} : \theta_i = 0$ is the union of all $H_{0I}^*$ with $i \in I$, the rejection of all $H_{0I}^*$ with $i \in I$ implies $\theta_i \neq 0$.

Closed testing can now be justified by noting:

- A level-$\alpha$ test for $H_{0I}$ is automatically a level-$\alpha$ test for $H_{0I}^*$;
- The union of all $H_{0I}$ with $i \in I$ is the same as the union of all $H_{0I}^*$ with $i \in I$.

Closed testing and partition testing are general principles of multiple testing. Hence, to test each $H_{0I}$ and $H_{0I}^*$, any level-$\alpha$ test can be used. For testing a set of null hypotheses $H_{0I}, i = 1, \ldots, k$ closed and partition testing methods are more powerful than the corresponding one-step multiple comparison methods (such as Tukey's, Scheffé's, and Dunnett's methods) because, in effect, one-step methods adjust for a multiplicity of $k$ in testing $H_{0I}$ or $H_{0I}^*$ even when $I \subset \{1, \ldots, k\}$. Also, partition testing is always at least as powerful as closed testing (due to the second justification, above, for closed testing). Finner and Strassburger (2002) showed

that, because partition testing allows a bigger class of tests than closed testing, it can be more powerful than closed testing in some applications. But in most applications closed testing and partition testing provide the same inference. Our subsequent discussion is mainly in terms of partition testing.

With suitable modeling of the observations, the joint distribution of estimates of differential expressions and their standard errors can be obtained. Thus, in principle, one can use this joint distribution to construct a level-$\alpha$ test for each $H_{0I}$ or $H^*_{0I}$ (without assuming independence or relying on probabilistic inequalities) and execute a closed or partition test. However, a direct implementation of closed testing or partition testing requires testing and collating the results of $2^g - 1$ null hypotheses, clearly a computational impossibility if the number of genes $g$ is more than a handful. Fortunately, appropriate choices of test statistics allow the testing of the vast majority of the hypotheses $H^*_{0I}$ (or $H_{0I}$) to be skipped, resulting in a stepdown test. Without such a shortcut, it would be practically impossible to implement any multiple testing method that gives strong control of the familywise error rate. In the following section, we describe conditions under which a closed or partition test can be executed as a stepdown test. Although these conditions have been well researched and applied in clinical biostatistics over the past decade, their importance has yet to be sufficiently appreciated in the statistical analysis of microarray data.

## 5 Stepdown Testing—A Shortcut to Closed and Partition Testing

The key condition needed to effect a shortcut is roughly that the rejection of a more restrictive hypothesis implies the rejection of certain less restrictive null hypotheses. So if one starts by testing the more restrictive null hypothesis and then skips the testing of less restrictive hypotheses as such implications allow, then closed or partition testing becomes more computationally feasible. The resulting shortcut version of a closed or partition test is a stepdown test.

Specifically, if gene $i_0$ is the gene that appears the most significantly differentially expressed among the genes whose indices are in the set $I$, and the rejection of the null hypothesis $H_{0I} : \{\theta_i = 0, \text{ for all } i \in I\}$ guarantees rejection of all null hypotheses which state that gene $i_0$ is not differentially expressed (possibly among others), then a shortcut can be taken. This would not be true if, for example, the test for $H_{0I} : \{\theta_i = 0, \text{ for all } i \in I\}$ is in the form of a $\chi^2$ or $F$-test, rejecting $H_{0I}$ if $\sum_{i \in I} \theta_i^2$ is large. Hsu (1996, pages 136–137) provided an example of an erroneous previous shortcutting computer implementation of such a statistical method, demonstrating with data explicitly why such shortcuts cannot be taken.

A precise set of sufficient conditions for such shortcutting to be valid is as follows.

S1: Tests for all hypotheses are based on statistics $T_i$, $i = 1, \ldots, g$, whose values do not vary with the null hypothesies $H^*_{0I}$ being tested;

S2: The level-$\alpha$ test for $H_{0I}^*$ is of the form of rejecting $H_{0I}^*$ if $\max_{i\in I} T_i > c_I$;
S3: Critical values $c_I$ have the property that if $J \subset I$ then $c_J \leq c_I$.

For example, suppose that g = 3 and the values of the statistics are $T_1 < T_3 < T_2$. If $|T_2| > c_{\{1,2,3\}}$ so that $H_{\{1,2,3\}}$ is rejected, then one does not need to test $H_{\{1,2\}}$ and $H_{\{2,3\}}$ because, by conditions S1 to S3, it is known that the results would be rejections. Thus, if conditions S1 to S3 are satisfied, then partition testing has the following shortcut. Let $[1], \ldots, [g]$ be the indices such that $T_{[1]} < \cdots < T_{[g]}$, then

Step 1: If $T_{[g]} > c_{\{[1],\ldots,[g]\}}$ then infer $\theta_{[g]} \neq 0$ and go to step 2; else stop;
Step 2: If $T_{[g-1]} > c_{\{[1],\ldots,[g-1]\}}$, then infer $\theta_{[g-1]} \neq 0$ and go to step 3; else stop;
...
Step g: If $T_{[1]} > c_{\{[1]\}}$, then infer $\theta_{\{[1]\}} \neq 0$ and stop.

A method of this form is called a stepdown test because it steps down from the most statistically significant to the least statistically significant.

Subtleties in conditions S1 to S3 for shortcutting have not always been appreciated. For example, suppose the critical values $c_1$ are computed so that the probability of $\max_{i\in I} T_i > c_I$ is less than or equal to $\alpha$ when $H_{\{1,\ldots,g\}}^* : \theta_1 = \cdots = \theta_g = 0$ is true, then the test for $H_I^*$ may or may not be a level-$\alpha$ test. This is because the distribution of $T_i$ for $i \in I$ may depend on the values of $\theta_j$, $j \notin I$. A level-$\alpha$ test for $H_I^*$ should satisfy

$$\sup_{\theta_{\in\Theta_I}} P_\theta\{\max_{i\in I} T_i > c_I\} \leq \alpha,$$

where $\Theta_I = \{\theta : \theta_i = 0 \text{ for } i \in I \text{ and } \theta_j \neq 0 \text{ for } j \notin I\}$ and the supremum of this rejection probability may or may not occur at $\theta_1 = \cdots = \theta_g = 0$. Conditions S1 to S3 guarantee that this undesirable phenomenon does not occur, thereby ensuring strong control of the FWER.

Another condition, similar to conditions S1 to S3, which also guarantees that a stepdown method strongly controls the familywise error rate is the *subset pivotality* condition proposed by Westfall and Young (1993, page 42). Their original subset pivotality condition is given in terms of adjusted $p$-values. For comparability with S1 to S3, we have paraphrased that condition here in terms of test statistics:

Subset pivotality: For all $I \subset \{1, \ldots, g\}$, the joint distribution of $(T_i, i \in I)$ under $\theta_i = 0, i \in I$, remains the same regardless of the values of $\theta_j$, $j \notin I$.

One should verify either conditions S1 to S3 are satisfied, or subset pivotality is satisfied, before implementing a stepdown test for, otherwise, the stepdown test may not strongly control the familywise error rate. Such conditions are easier to check with a model that connects the observations with the parameters, but harder to check with a model (such as the randomization model) that only describes the distribution of the observations under the null hypotheses. Indeed, Westfall and Young (1993, page 91) cautioned that the randomization model does not guarantee that the subset pivotality condition holds. Outside the context of bioinformatics, there are in fact examples of methods that were in use at one time that violate

these conditions (see, for example, Hsu, 1996, Chapter 6). In the context of gene expressions, if it is possible for the correlations between nondifferentially expressed genes and differentially expressed genes to be different for normal and disease tissues, then it seems to us that perhaps not all permutations of observations from genes that are not differentially expressed are equally likely across normal and disease conditions. In that case, subset pivotality does not hold, and conditions S1 to S3 may not hold, so stepdown permutation tests may not be valid. For this reason, we believe multiple testing of gene expressions without modeling requires careful validation.

## 6 An Example

Measurements of gene expression levels not only depend on true gene expression under the treatments or conditions being compared, but potentially also on array and dye variations. The amount of cDNA deposited may differ from one array to the next. Genes might have differential affinity for the Cy3 (green) dye or the Cy5 (red) dye. The nuisance factors "array" and "dye" should be treated as blocking factors in designing microarray experiments in order to eliminate their effects on the analysis of the gene expressions.

For comparing two "treatments" $t_1$ and $t_2$ (such as tissue types or cell lines), both Kerr et al. (2000) and Hsu et al. (2002) recommended the design of two-channel microarray experiments as generalized Latin squares with an even number $a$ of arrays and two dyes, as follows.

1. Randomly assign treatment labels $t_1$ and $t_2$ to treatments (conditions) 1 and 2;
2. Randomly assign array labels $1, \ldots, a$ to the actual microarrays;
3. Randomly assign dye labels 1 and 2 to the Cy5 (red) and Cy3 (green) dyes;
4. Assign treatment labels $t_1$ and $t_2$ to dyes and arrays according to Table 1;
5. Randomly spot each gene $n$ times on each array.

This is an example of what are called "dye-swap experiments" by Churchill (2002) and by Yang and Speed (2002).

To illustrate the modeling and subsequent analysis of gene expression levels, we use the Synteni data of Kerr et al. (2000) which compared the expressions of 1286 genes of a human liver tissue sample to the same genes of a muscle tissue sample. So, in this example, the two treatments are the two tissue types. The microarray experiment was designed according to a generalized Latin square design as in Table 1 with each of the 1286 genes spotted once on each array. The data can be modeled successfully by a linear model, described below. Let $y_{mdkir}$ be the observed gene expression of the $m$th array, $d$th dye, $k$th treatment, $i$th gene, and $r$th

TABLE 1. A generalized Latin square of size 2 × a

| Dye/Array | 1 | 2 | 3 | ... | $a-1$ | $a$ |
|---|---|---|---|---|---|---|
| 1 | $t_1$ | $t_2$ | $t_1$ | ... | $t_1$ | $t_2$ |
| 2 | $t_2$ | $t_1$ | $t_2$ | ... | $t_2$ | $t_1$ |

replication, and let $y'_{mdkir} = \log_e(y_{mdkir})$. Then a possible model for the logarithm of observed gene expression levels is

$$y'_{mdkir} = \mu + \alpha_m + \delta_d + \tau_k + \gamma_i + (\alpha\gamma)_{mi} + (\delta\gamma)_{di} + (\tau\gamma)_{ki} + \varepsilon_{mdkir}, \quad (2)$$
$$m = 1, \ldots, a;\ d = 1, 2;\ k = 1, 2;\ i = 1, 2, \ldots, g;\ r = 1, 2, \ldots, n;$$

where $\mu$ is the overall mean, $\alpha_m$ is the effect of array $m$, $\delta_d$ is the effect of dye $d$, $\tau_k$ is the effect of treatment $k$, $\gamma_i$ is the effect of gene $i$, $(\alpha\gamma)_{mi}$ is the interaction effect of array $m$ and gene $i$, $(\delta\gamma)_{di}$ is the interaction of dye $d$ and gene $i$, $(\tau\gamma)_{ki}$ is the interaction effect of treatment $k$ and gene $i$, and the errors $\varepsilon_{mdkir}$ are independent $N(0, \sigma^2)$. Inclusion of the array and dye effects in the model "pre-processes" the data, removing from the analysis the effects of the nuisance factors and dependence among the observations caused by such factors. In the above experiment, $n = 1$.

After deleting the outlying observations from one particular gene, regression diagnostics show that, for the remaining 1285 genes, the reduced model

$$y'_{mdkir} = \mu + \alpha_m + \delta_d + \tau_k + \gamma_i + (\delta\gamma)_{di} + (\tau\gamma)_{ki} + \varepsilon_{mdkir}, \quad (3)$$

with independent normally distributed errors having equal variances, is adequate.

As Kerr et al. (2000) gave no indication that housekeeping genes were included in the microarray experiment, the average expression level (averaged over arrays and dyes) of all the genes serves as the (negative) control for normalizing the data. The parameters of interest are, thus, the mean differences of log intensities between treatments for the gene, compared to the mean difference in log intensities between treatments averaged over all genes:

$$\begin{aligned}\theta_i &= \tau_1 + \gamma_i + (\tau\gamma)_{1i} - (\tau_2 + \gamma_i + (\tau\gamma)_{2j}) \\ &\quad - \left[\tau_1 + \frac{\sum_{j=1}^{g}\gamma_j}{g} + \frac{\sum_{j=1}^{g}(\tau\gamma)_{1j}}{g} - \left(\tau_2 + \frac{\sum_{j=1}^{g}\gamma_j}{g} + \frac{\sum_{j=1}^{g}(\tau\gamma)_{2j}}{g}\right)\right] \\ &= (\tau\gamma)_{1i} - (\tau\gamma)_{2i} - [\overline{\tau\gamma}_{1\cdot} - \overline{\tau\gamma}_{2\cdot}],\end{aligned}$$

for $i = 1, \ldots, g$, where a dot in the subscript and an overbar indicate averaging over the levels of the corresponding factor. The least squares estimators of $\theta_i (i = 1, 2, \ldots, g)$ are

$$\hat{\theta}_i = \bar{y}'_{\cdot\cdot 1\cdot} - \bar{y}'_{\cdot\cdot 2i\cdot} - \left[\frac{\sum_{j=1}^{g}\bar{y}'_{\cdot\cdot 1j\cdot}}{g} - \frac{\sum_{j=1}^{g}\bar{y}'_{\cdot\cdot 2j\cdot}}{g}\right].$$

The estimators have a (singular) multivariate normal distribution, with means $\theta_i, i = 1, \ldots, g$, equal variances, and equal correlations $-1/(g-1)$. Let $SE(\hat{\theta}_i)$ denote the estimate of the standard error of $\hat{\theta}_i$, which is independent of $i$ and hence common to all genes.

Whereas Hsu et al. (2002) obtained simultaneous confidence intervals for $\theta_i$, $i = 1, \ldots, g$, here we illustrate partition testing of the null hypotheses

$$H^{\star}_{0I} : \{\theta_i = 0 \text{ for all } i \in I \text{ and } \theta_j \neq 0 \text{ for all } j \notin I\}. \tag{4}$$

We choose the following level-$\alpha$ test for (4). We reject $H^{*}_{0I}$ if

$$\max_{i \in I} |T_i| = \max_{i \in I} \left| \frac{\hat{\theta}_i - \theta_i}{SE(\hat{\theta}_i)} \right| > c_I, \tag{5}$$

where the critical value $c_I$ is the upper $100(1-\alpha)$ percentile of the distribution of $\max_{i \in I} |\hat{\theta}_i - \theta_i| / SE(\hat{\theta}_i)$. Because $\theta_i (i = 1, \ldots, g)$ are equi-correlated, $c_I$ depends only on the cardinality $|I|$ of $I$. So let us denote $c_I$ by $c_{|I|}$. The computation of the exact value of $c_{|I|}$ is possible but, if the number of genes $g$ is large, then $c_{|I|}$ is well approximated by the $100(1-\alpha)$ percentile of the Studentized maximum modulus distribution with $|I|$ treatments and the error degrees of freedom given by the analysis of variance table. The reason is that, for the Studentized maximum modulus distribution to be valid, the estimators $\hat{\theta}_i - \theta_i, i \in I$, need to be independent for any $I \subseteq \{1, \ldots, g\}$, and this is approximately true if $g$ is large (see Section 1.3.1 of Hsu, 1996, and also Fritsch and Hsu, 1997.)

Without shortcutting, closed or partition testing would require testing $2^{1285} - 1$ tests, which is clearly impossible to achieve. However, conditions S1 and S2 for shortcutting closed and partition testing are trivially satisfied. Also, condition S3 is satisfied because the distribution of $\max_{i \in I} |T_i|$ does not depend on any parameter, including $\theta_j$, $j \notin I$.

For this data set, at a familywise error rate of 5%, one-step testing leads to the conclusion that 174 genes are differentially expressed. Closed and partition stepdown testing are guaranteed to infer at least those same 174 genes to be differentially expressed. In fact, these last two methods both infer three additional genes to be differentially expressed; that is, a total of 177 genes.

To guard against the possibility of nonnormal errors, we also applied the bootstrap-$t$ technique (see Beran, 1988, and Efron and Tibshirani, 1993); that is, we bootstrapped the residuals to estimate the quantiles of $\max_{i \in I} |T_i|$, and used them as critical values for testing. At a FWER of 5%, one-step bootstrap testing infers 161 genes to be differentially expressed, whereas stepdown bootstrap testing infers 164 genes, with three additional genes, to be differentially expressed.

Figure 1 shows plots of the ordered $|T_i|$ values (solid dots) with normal distribution critical values (dotted line) and bootstrap critical values (dashed line). On the horizontal axis is the gene index that sorts the genes so that 1 corresponds to the gene with the smallest $|T|$ value, and 1285 corresponds to the gene with the largest $|T|$ value. A stepdown method starts by checking whether the largest $|T|$ value is larger than the largest critical value. If it is, then the gene corresponding to that largest $|T|$ value is inferred to be differentially expressed. Then one checks whether the second largest $|T|$ value is larger than the second largest critical value,

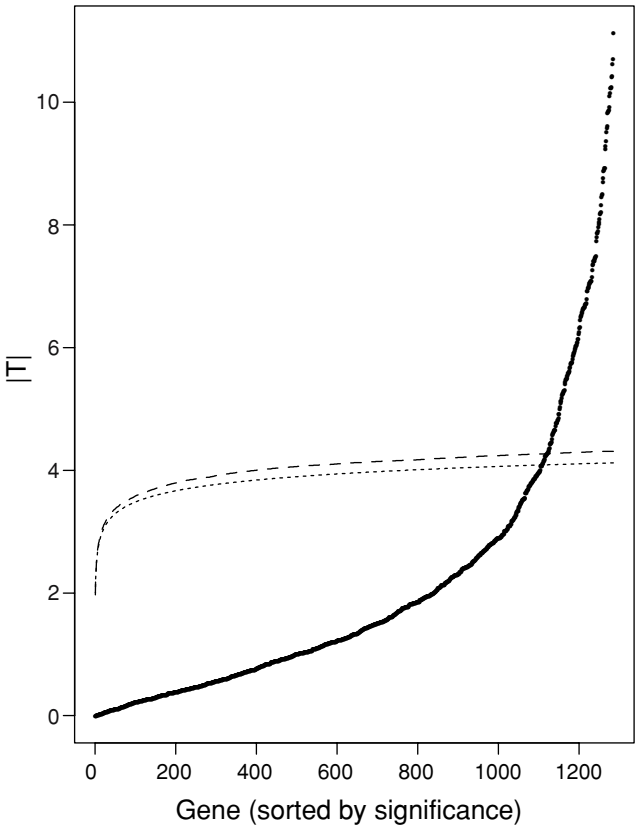



FIGURE 1. Ordered $|T_i|$ values (solid dots) with normal distribution critical values (dotted line) and bootstrap critical values (dashed line) for the stepdown method.

and so on. The stepdown method stops as soon as an ordered $|T|$ value is less than the corresponding critical value. Thus, starting from the right of Figure 1 and moving towards the left, those genes with solid dots above the critical value line are inferred to be differentially expressed.

## 7 Discussion

For the potential of gene expression analysis to be fully realized, we believe good choices of statistical design and analysis methods in microarray experiments are essential. A good design not only avoids bias in estimation from nuisance factors, but also allows the analysis to be executed efficiently. A good analysis method guards against finding an excessive number of false positives and, hence, maximizes the probability of making a truly useful scientific discovery.

*Acknowledgments.* We thank Gary Churchill, Ramana Davuluri, Soledad Fernandez, Marilisa Gibellato, Gene J.T. Hwang, Kenton Juhlin, Yoonkyung Lee, Steven MacEachen, James Rogers, A. A. Scott, and Gunnar Stefansson for insightful discussions.

## *References*

Alberts, B., Johnson, A., Lewis, J., Raff, M, Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*, fourth edition. Garland, New York.

Beran, R. (1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, **83**, 679–686.

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, **32**, 490–495.

Churchill, G. A. (2003). Discussion of "Statistical challenges in functional genomics." *Statistical Science*, **18**, 64–69.

Churchill, G. A. and Oliver, B. (2001). Sex, flies and microarrays. *Nature Genetics*, **29**, 355–356.

Dudoit, S. J., Shaffer, P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

Dudoit, S., Yang, Y. H., Speed, T. P., and Callow, M. J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151–1161.

Finner, H. and Roter, M. (2001). On the false discovery rate and expected type I errors. *Biometrical Journal*, **43**, 985–1005.

Finner, H. and Strassburger, K. (2002). The partitioning principle: A powerful tool in multiple decision theory. *Annals of Statistics*, **30**, 1194–1213.

Fritsch, K. and Hsu, J. C. (1997). On analysis of means. In *Advances in Statistical Decision Theory and Methodology*. Editors: N. Balakrishnan and S. Panchapakesan. Birkhäuser, Boston, 114–119.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley and Sons, New York.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall, London.

Hsu, J. C., Chang, J. Y., and Wang, T. (2002). Simultaneous confidence intervals for differential gene expressions. *Technical Report 592*, The Ohio State University, Columbus.

ICH E10 (1999). *Choice of Control Groups in Clinical Trials*. CPMP (Committee for Propritary Medical Products), EMEA (The European Agency for the Evaluation of Medical Products), London, Draft ICH (International Conference on Harmonisation). Efficiency guidelines, http://www.ich.org.

Kerr, M. K. and Churchill, G. (2001a). Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Kerr, M. K. and Churchill, G. (2001b). Statistical design and the analysis of gene expression microarray data. *Genetical Research*, **77**, 123–128.

Kerr, M. K., Martin, M., and Churchill, G. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, **7**, 819–837.

Lee, M. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the U.S.A.*, **18**, 9834–9839.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

Rao, C. R. (1973). Linear Statistical Inference and Its applications, second edition, John Wiley and Sons, New York.

Stefansson, G., Kim, W., and Hsu, J. C. (1988). On confidence sets in multiple comparisons. In *Statistical Decision Theory and Related Topics IV*, volume 2. Editors: S. S. Gupta and J. O. Berger, pages 89–104. Springer-Verlag, New York.

Thomas, J. G., Olson, J. M., Tapscott, S. J., and Zhao, L. P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, **11**, 1227–1236.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the U.S.A.*, **98**, 5116–5121.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, **3**, 0034.I–0034.II.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley and Sons, New York.

Wolfinger, R., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, **8**, 625–637.

Yang, Y. H. and Speed, T. P. (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, **3**, 579–588.

# 7
# Projection Properties of Factorial Designs for Factor Screening

Ching-Shui Cheng

The role of projection in screening is discussed and a review of projection properties of factorial designs is provided. The "projection of a factorial design onto a subset of factors" is the subdesign consisting of the given subset of factors (or, equivalently, the subdesign obtained by deleting the complementary set of factors). A factor-screening design with good projections onto small subsets of factors can provide useful information when a small number of active factors have been identified. The emphasis in this chapter is placed on projection properties of nonregular designs with complex aliasing. The use of projection in search designs and uniform designs is also discussed briefly.

## 1 Introduction

In the initial stage of experimentation, there may be a large number of potentially important factors, however, according to the principle of effect sparsity (Box and Meyer, 1986), often only a few of these factors are "active". We say that a factor is *active* if it is involved in at least one nonnegligible factorial effect. Thus, in designing experiments for factor screening, it is important to consider properties of the subdesigns consisting of small subsets of factors, called *projections* or *projection designs*. A factor-screening design with good projections onto small subsets of factors provides useful information after the small number of active factors has been identified. Traditionally, designs with complicated alias structures, such as Plackett–Burman designs, have been used for studying main effects, under the assumption that all of the interactions are negligible. Interactions are typically ignored due to the difficulty in disentangling complicated aliasing among the factorial effects. However, if there are only a small number of active factors, then such aliasing may no longer be a problem and it may be possible to study some interactions. Hamada and Wu (1992) proposed a strategy to entertain and estimate two-factor interactions from Plackett–Burman type designs. In their work on screening and prediction in computer experiments, Welch et al. (1992) advocated using the data from a small experiment not only to screen factors but also to build an accurate predictor when the computer code is computationally expensive, instead of treating the task of screening as a separate preliminary activity with

follow-up experiments. More recently, Cheng and Wu (2001) also proposed an approach that performs screening and response surface exploration using a single design: the first-stage design is projected onto the factors that are identified as important and then response surface exploration is conducted using the projection design. The key to success lies in good low-dimensional projection properties of the initial design. In this context, it is necessary to consider projection onto every small subset of factors inasmuch as it is not known a priori which factors are active.

This chapter provides a review of some recent work on projection properties of factorial designs. In Section 2, the concepts of orthogonal arrays and projectivity are reviewed and a detailed example is provided to illustrate the concept of projection. In Section 3, projection properties of regular designs are summarized, and Section 4 is devoted to nonregular designs. Projection and search designs are discussed in Section 5. A brief review of the use of projection in uniform and space-filling designs, which arise in numerical integration, computer experiments, and drug discovery, is presented in Section 6. The chapter ends with some concluding remarks.

## 2 Orthogonal Arrays and Projectivity

Orthogonal arrays, introduced by C. R. Rao (1947), have been used extensively in factorial designs. An orthogonal array, denoted OA$(n, s^f, t)$, is an $n \times f$ matrix $\boldsymbol{D}$ of s symbols such that all the ordered $t$-tuples of the symbols occur equally often as row vectors of any $n \times t$ submatrix of $\boldsymbol{D}$. The array is said to be of *strength* $t$. Each OA $(n, s^f, t)$ defines an $n$-run factorial design for $f$ factors, each having $s$ levels, where the symbols represent factor levels, columns correspond to factors, and rows represent factor-level combinations. In general, an orthogonal array of strength $t = 2l$ can be used to estimate all the main effects and interactions involving at most $l$ factors under the assumption that all the interactions involving more than $l$ factors are negligible. On the other hand, an orthogonal array of strength $t = 2l - 1$ can be used to estimate all the main effects and interactions involving at most $l - 1$ factors under the assumption that all the interactions involving more than $l$ factors are negligible. For example, an orthogonal array of strength two defines a design under which all the main effects can be estimated when the interactions are negligible; see Hedayat et al. (1999) for more technical details of orthogonal arrays.

Implicit in the definition of an orthogonal array is the following important projection property: in the projection of an orthogonal array of strength $t$ onto any subset of $t$ factors, all the factor-level combinations are replicated the same number of times. Box and Tyssedal (1996) defined a design to be of *projectivity* $p$ if, for every subset of $p$ factors, a complete factorial design (possibly with some combinations replicated) is produced. This property can be viewed as an extension of the concept of strength. It was also discussed by Constantine (1987, page 363) and has further appeared in another context: $p$-projectivity is the same as $p$-covering in designs for circuit testing; see, for example, Sloane (1993). If a design has projec-

TABLE 1. An example of an OA(16, $2^6$, 3)

| Factors | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| −1 | −1 | −1 | −1 | −1 | −1 |
| −1 | −1 | −1 | 1 | −1 | 1 |
| −1 | −1 | 1 | −1 | 1 | 1 |
| −1 | −1 | 1 | 1 | 1 | −1 |
| −1 | 1 | −1 | −1 | 1 | −1 |
| −1 | 1 | −1 | 1 | 1 | 1 |
| −1 | 1 | 1 | −1 | −1 | 1 |
| −1 | 1 | 1 | 1 | −1 | −1 |
| 1 | −1 | −1 | −1 | 1 | 1 |
| 1 | −1 | −1 | 1 | 1 | −1 |
| 1 | −1 | 1 | −1 | −1 | −1 |
| 1 | −1 | 1 | 1 | −1 | 1 |
| 1 | 1 | −1 | −1 | −1 | 1 |
| 1 | 1 | −1 | 1 | −1 | −1 |
| 1 | 1 | 1 | −1 | 1 | −1 |
| 1 | 1 | 1 | 1 | 1 | 1 |

tivity $p$ and no more than $p$ active factors, then no matter which factors are active, all the main effects and interactions of the active factors can be estimated when the design is projected onto the active factors.

Table 1 displays an orthogonal array of strength three because it contains each of the $2^3 = 8$ level combinations of any set of three factors exactly twice. In other words, its projection onto any set of three factors consists of two replicates of the complete $2^3$ factorial design.

An orthogonal array is called *regular* if it can be constructed by using a "defining relation". For example, the array displayed in Table 1 is regular and is constructed by the following method. Denote a row (factor-level combination) of the array by $(x_1, \ldots, x_6)$. The first four columns of the array consist of all the 16 level combinations of factors 1 through 4, and the levels of factors 5 and 6 are defined via

$$x_5 = x_1x_2x_3 \quad \text{and} \quad x_6 = x_1x_3x_4.$$

Therefore, all the 16 factor-level combinations in the array satisfy

$$x_1x_2x_3x_5 = x_1x_3x_4x_6 = 1$$

and so the product

$$(x_1x_2x_3x_5)(x_1x_3x_4x_6) = x_2x_4x_5x_6$$

is also equal to 1. Thus, the array consists of solutions of the *defining equations*

$$1 = x_1x_2x_3x_5 = x_1x_3x_4x_6 = x_2x_4x_5x_6.$$

This is often written as $I = 1235 = 1346 = 2456$, and is called the *defining relation* of the design. Each term in the defining relation is called a *defining word*. If we multiply each defining word by 1, then we obtain $1 = 235 = 346 = 12456$. This means that when only the 16 factor-level combinations in the array of Table 1 are observed, the main effect of factor 1, the interaction of factors 2, 3, 5, the interaction of factors 3, 4, 6, and the interaction of factors 1, 2, 4, 5, and 6 are completely mixed up; we say that they are *aliases* of one another. The alias relations of factorial effects under a regular design can be obtained easily from its defining relation. The reader is referred to Chapter 1 of this book or Wu and Hamada (2000, Chapter 4) for more detailed discussions of regular designs.

The length of the shortest defining word of a regular design is called its *resolution*. A regular design of resolution $t$ is an orthogonal array of strength $t - 1$. For example, the array displayed in Table 1 is a design of resolution four.

## 3 Projection Properties of Regular Designs

Projection properties of a regular design can easily be determined from its defining relation. We refer the readers to Chen (1998) for various results on projection properties of regular designs. The projection of a regular design onto any subset of factors is also regular, either a regular fractional factorial design or a set of replicates of a regular fractional factorial design. All defining equations satisfied by the factor-level combinations in a projection design must also be satisfied by the factor-level combinations in the parent design. Thus the defining words of the projection design are precisely those defining words of the parent design that involve only factors in the projection design.

For example, consider the $2^{8-4}$ resolution IV design defined by

$$I = 1235 = 1346 = 2347 = 1248 = 2456 = 1457 = 3458 = 1267$$
$$= 2368 = 1378 = 3567 = 1568 = 2578 = 4678 = 12345678.$$

Among the 15 defining words, exactly three (3458, 3567, and 4678) involve factors 3, 4, 5, 6, 7, and 8 only. Therefore the projection onto these six factors is the $2^{6-2}$ design defined by $I = 3458 = 3567 = 4678$. Because no defining word of the above $2^{8-4}$ design involves factors 1, 2, 3, and 4 only, the projection onto these four factors is a $2^4$ complete factorial design. On the other hand, one defining word of the above design involves factors 1, 2, 3, and 5; therefore the projection onto these factors consists of two replicates of the $2^{4-1}$ design defined by $I = 1235$. Such a projection does not contain all the 16 factor-level combinations. Thus the above $2^{8-4}$ design does not have projectivity four but, because it contains all eight level combinations of any set of three factors, it has projectivity three.

In general, if a regular design has resolution $t$ (maximal strength $t - 1$), then it is of projectivity $t - 1$, but cannot be of projectivity $t$. This is because the projection of the design onto the factors that appear in a defining word of length $t$ is a replicated $2^{t-1}$ fractional factorial design. Contrary to this, we shall see in

the next section that a nonregular orthogonal array of maximal strength $t-1$ *can* have projectivity $t$.

## 4 Projection Properties of Nonregular Designs

When a regular design is used, any two factorial effects (main effects or interactions) are either orthogonal or are completely aliased. Partial aliasing of factorial effects makes the alias structures of nonregular designs much more complex (see Chapter 8 for information on the difficulties in the analysis when complex aliasing is present), but it also leads to some superior projection properties.

Plackett–Burman designs are important examples of nonregular orthogonal arrays of strength two, and are constructed from Hadamard matrices (see, for example, Wu and Hamada, 2000, Chapter 7). The economy in run sizes provided by such designs makes them suitable for screening experiments. A *Hadamard matrix* of order $n$ is an $n \times n$ matrix with entries $+1$ and $-1$ in which any two columns (and hence any two rows) are orthogonal in the usual sense of having a zero inner product. Such a matrix can be "normalized" by multiplying all the entries of certain rows by $-1$ so that all the entries in the first column become equal to 1. An OA$(n, 2^{n-1}, 2)$ can then be obtained by deleting the first column. Such a design can accommodate $n-1$ two-level factors in $n$ runs and can be used to estimate all the main effects when the interactions are assumed to be negligible.

Table 2 shows a 12-run Plackett–Burman design. All OA$(12, 2^{11}, 2)$ are equivalent in the sense that they can be obtained from one another by changing signs in one or more complete columns and/or permuting rows and/or permuting columns. Hadamard matrices of higher orders, however, are not unique.

Projection properties of Plackett–Burman designs were studied by Lin and Draper (1991, 1992). Their computer searches examined all of the projections of small Plackett–Burman designs onto a few factors. For example, each of the 165 projections of the 12-run design of Table 2 onto three factors consists of

TABLE 2. An example of an OA$(12, 2^{11}, 2)$

| Factors | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| −1 | −1 | 1 | −1 | −1 | −1 | 1 | 1 | 1 | −1 | 1 |
| 1 | −1 | −1 | 1 | −1 | −1 | −1 | 1 | 1 | 1 | −1 |
| −1 | 1 | −1 | −1 | 1 | −1 | −1 | −1 | 1 | 1 | 1 |
| 1 | −1 | 1 | −1 | −1 | 1 | −1 | −1 | −1 | 1 | 1 |
| 1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 | −1 | −1 | 1 |
| 1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 | −1 | −1 |
| −1 | 1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 | −1 |
| −1 | −1 | 1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 |
| −1 | −1 | −1 | 1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 |
| 1 | −1 | −1 | −1 | 1 | 1 | 1 | −1 | 1 | −1 | −1 |
| −1 | 1 | −1 | −1 | −1 | 1 | 1 | 1 | −1 | 1 | −1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

a complete $2^3$ design and a half-replicate of a $2^3$ design. Thus, it is of projectivity three. This property was also observed by Box and Bisgaard (1993), who commented that the interesting projective properties of Plackett–Burman designs, which experimenters have sometimes been reluctant to use for industrial experimentation due to their complicated alias structures, provide a compelling rationale for their use. This superior projection property of the 12-run Plackett–Burman design is not shared by regular designs, as we have seen in the previous section.

It is the existence of a defining word of length $t$ that prevents a regular design of resolution $t$ from having projectivity $t$. This key concept can be extended to nonregular designs. For simplicity, this extension is described here for only two-level designs.

Suppose an $n$-run design with $f$ two-level factors is represented by an $n \times f$ matrix $\boldsymbol{D}$ with elements $d_{ij}$, $(i = 1, \ldots, n;\ j = 1, \ldots, f)$, where $d_{ij}$ is 1 or $-1$ representing the level of the $j$th factor on the $i$th run. If we denote the $j$th column of $\boldsymbol{D}$ by $\boldsymbol{d}_j$, then the elements of $\boldsymbol{d}_j$ define a contrast representing the main effect of the $j$th factor. For each $2 \le l \le f$ and any subset $S = \{j_1, \ldots, j_l\}$ of $\{1, \ldots, f\}$, let

$$\boldsymbol{d}_S = \boldsymbol{d}_{j1} \odot \cdots \odot \boldsymbol{d}_{jl},$$

where, for $x = (x_1, \ldots, x_n)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{x} \odot \boldsymbol{y}$ is their Hadamard, or elementwise, product; that is,

$$\boldsymbol{x} \odot \boldsymbol{y} = (x_1 y_1, \ldots, x_n y_n)^T,$$

where $T$ denotes transpose. Then $\boldsymbol{d}_S$ is a contrast corresponding to the interaction of the factors in $S$. The design is said to have a *defining word of length r* if there exist $r$ columns, say columns $j_1, \ldots, j_r$, such that $\boldsymbol{d}_{j1} \odot \cdots \odot \boldsymbol{d}_{jr} = \mathbf{1}_n$ or $-\mathbf{1}_n$, where $\mathbf{1}_n$ is the $n \times 1$ vector of 1s. Two factorial effects are said to be *completely aliased* if their corresponding columns, say $\boldsymbol{d}_{S_1}$ and $\boldsymbol{d}_{S_2}$, are such that

$$\text{either} \quad \boldsymbol{d}_{S_1} = \boldsymbol{d}_{S_2} \quad \text{or} \quad d_{S_1} = -\boldsymbol{d}_{S_2};$$

that is, $|\boldsymbol{d}_{S_1}^T \boldsymbol{d}_{S_2}| = n$. For example, we have already seen that the regular design in Table 1 has a defining word 1235. Indeed, the Hadamard product of columns 1, 2, 3, and 5 of this array is equal to the vector of all 1s. Because the Hadamard product of columns 1 and 2 is equal to that of columns 3 and 5, we see that the interaction of factors 1 and 2 is completely aliased with that of factors 3 and 5. In a regular design, we have

$$\text{either} \quad |\mathbf{1}_n^T \boldsymbol{d}_S| = n \quad \text{or} \quad \mathbf{1}_n^T \boldsymbol{d}_S = 0$$

for all subsets $S$ of $\{1, \ldots, f\}$. Thus any two factorial effects are either orthogonal or completely aliased. For nonregular designs, we could have $|\mathbf{1}_n^T \boldsymbol{d}_S|$ strictly between 0 and $n$, leading to partial aliasing. For example, for the 12-run design in Table 2, $|\mathbf{1}_n^T \boldsymbol{d}_S| = 4$ for all subsets $S$ of size three. Even though this design only has strength two, it does not have any defining word of length 3 or 4; that is, no $\boldsymbol{d}_S$ is $+\mathbf{1}_n$ or $-\mathbf{1}_n$ for subsets of size 3 or 4. This is not possible with regular designs.

It turns out that, in general, the nonexistence of defining words of length $t + 1$ in an OA$(n, 2^f, t)$ with $f \ge t + 1$ is necessary and sufficient for the design, regular

or not, to be of projectivity at least $t + 1$; notice that, because the array has strength $t$, it has no defining words of lengths shorter than $t + 1$. It is, in fact, straightforward to show that the condition is necessary and sufficient for regular designs and that it is necessary for nonregular designs. Cheng (1995) showed that the condition is also sufficient for nonregular designs. Thus, we have the following result.

**Theorem 1.** *An OA($n, 2^f, t$) with $f \geq t + 1$ has projectivity $t + 1$ if and only if it has no defining word of length $t + 1$.*

The reason why a 12-run Plackett–Burman design is of projectivity three is that it does not have a defining word of length three. The 12-run Plackett–Burman design, although of projectivity three, is not of projectivity four. This is because, in order to have projectivity four, a two-level design must have at least 16 runs. Nevertheless, Lin and Draper (1993) and Wang and Wu (1995) observed an interesting property of projections of the 12-run Plackett–Burman design onto any four factors. By computer enumeration, they found that every such projection allows the estimation of all the main effects and two-factor interactions, assuming that three-factor and four-factor interactions are negligible, although their estimators may not be independent. Wang and Wu (1995) coined the term *hidden projection*, and also studied such projection properties of some other small Plackett–Burman type designs. The success of Hamada and Wu's (1992) strategy for entertaining and estimating two-factor interactions from Plackett–Burman type designs was attributed to the hidden projection properties of these designs. The hidden projection properties provide further support for the use of such designs for factor screening under the assumption of effect sparsity, even when some interactions are present.

Like projectivity, the hidden projection property is also tied to the nonexistence of defining words of short lengths. It is clear that, if an orthogonal array of strength two (regular or nonregular) has at least one defining word of length three or four, then some two-factor interactions are completely aliased with main effects or other two-factor interactions. In this situation, there are subsets of four factors for which the projection design cannot estimate all main effects and two-factor interactions. Thus, the nonexistence of defining words of length three or four is necessary for the above four-factor hidden projection property to hold. For regular designs, it is clearly also sufficient. Cheng (1995) showed that the condition is necessary and sufficient in general for two-level orthogonal arrays of strength two. This leads to the next theorem.

**Theorem 2.** *A necessary and sufficient condition for an OA($n, 2^f, 2$) with $f \geq 4$ to have the property that the main effects and two-factor interactions can be estimated in all projections onto four factors is that it has no defining words of length three or four.*

Again, a 12-run Plackett–Burman design has the four-factor hidden projection property stated in Theorem 2 because it has no defining words of length three or four. On the other hand, it is not possible for a regular design to have the same property unless it is of resolution five.

Cheng (1998a) further proved the following result for two-level orthogonal arrays of strength three.

**Theorem 3.** *A necessary and sufficient condition for an OA($n, 2^f, 3$) with $f \geq 5$ to have the property that the main effects and two-factor interactions can be estimated in all projections onto five factors is that it has no defining words of length four.*

By simple counting, it can be seen that if $n$ is not a multiple of 8, then no OA($n, 2^f, 2$) with $f \geq 4$ can have any defining words of length three or four. This property, together with Theorems 1 and 2 above, lead to the results of Cheng (1995) that such a design has projectivity three and that, in all of its four-factor projections, the main effects and two-factor interactions can be estimated. All Plackett–Burman designs whose run sizes are not multiples of 8 are covered by this result. Similarly, if $n$ is not a multiple of 16, then no OA($n, 2^f, 3$) with $f \geq 5$ has a defining word of length four, and Theorem 3 is applicable. In particular, if $n$ is not a multiple of 8, then the "foldover" (see Chapter 1) of any OA($n, 2^f, 2$) with $f \geq 4$ has the property that, in all its five-factor projections, the main effects and two-factor interactions can be estimated. Again, this result covers foldovers of Plackett–Burman designs whose run sizes are not multiples of 8.

What happens if $n$ is a multiple of 8? Cheng (1995) showed that, if the widely held conjecture is true that a Hadamard matrix exists for every order that is a multiple of 4, then for every $n$ that is a multiple of 8, one can always construct an OA($n, 2^f, 2$) that has defining words of length three or four. Such designs are not of projectivity 3, a fact also pointed out by Box and Tyssedal (1996), and do not have the desirable hidden projection properties mentioned above. However, this does not mean that, when $n$ is a multiple of 8, there are no OA($n, 2^f, 2$)s with good hidden projection properties. One notable example is the family of Paley designs, described below.

Many of the Plackett–Burman designs are constructed by a method due to Paley (1933). Suppose that $n$ is a multiple of 4 such that $n - 1$ is a power of an odd prime number (for example, $n = 12$ or 20, and so on). Let $q = n - 1$ and let $\alpha_1 (= 0)$, $\alpha_2, \ldots, \alpha_q$ denote the elements of *GF(q)*, the Galois field with $q$ elements; see Dey and Mukerjee (1999, Section 3.4) for a review of Galois fields. Define a function $\chi$ that maps the elements of *GF(q)* to the values 1 and $-1$ as follows,

$$\chi(\beta) = \begin{cases} 1, \text{ if } \beta = y^2 \text{ for } y \in GF(q), \\ -1, \text{ otherwise.} \end{cases}$$

Let $\boldsymbol{A}$ be the $q \times q$ matrix with $(i, j)$th element $a_{ij}$, where $a_{ij} = \chi(\alpha_i - \alpha_j)$ for $i, j = 1, 2, \ldots, q$, and define

$$\boldsymbol{P}_n = \begin{bmatrix} -\mathbf{1}_q^T \\ \mathbf{A} \end{bmatrix}.$$

Then $\boldsymbol{P}_n$ is an OA($n, 2^{n-1}, 2$), and is called a *Paley design*. For $n = 12$, the Paley design is the 12-run Plackett–Burman design. Paley designs can be constructed, for example, for $n = 12, 20, 24, 28, 32, 44$. By using a result from number theory, Bulutoglu and Cheng (2003) showed that all Paley designs with $n \geq 12$ have no defining words of length three or four. Therefore they all have projectivity three, the main effects and two-factor interactions can be estimated in all four-factor projections, and the same hidden projection property holds for all five-factor projections of their foldover designs. This result covers 24- and 32-run

Plackett–Burman designs even though 24 and 32 are multiples of 8. In fact, it can be verified directly that the 32-run Plackett–Burman design has the stronger property that all the main effects and two-factor interactions can be estimated in all its six-factor projections (Cheng, 1998b).

Projection properties of designs with more than two levels have been relatively unexplored. Cheng and Wu (2001) found a nonregular OA(27, $3^8$, 2) such that the quadratic model can be estimated in all four-factor projections. For $f$ factors, each with three quantitative levels, the quadratic model specifies the mean response at factor level combination $(x_1, \ldots, x_f)$ to be

$$\mu + \sum_{i=1}^{f} \beta_i x_i + \sum_{i=1}^{f} \beta_{ii} x_i^2 + \sum_{1 \le i < j \le f} \beta_{ij} x_i x_j,$$

where $\mu, \beta_i, \beta_{ii}, \beta_{ij}$ are unknown constants. The four-factor projection property mentioned above cannot be satisfied by any regular $3^{8-5}$ design (with 27 runs), again due to the presence of some defining words of length three. The design of Cheng and Wu (2001) is constructed by taking the union of three carefully chosen disjoint regular $3^{8-6}$ fractional factorial designs of resolution two. Cheng and Wu's design almost achieves the same hidden projection property for five-factor projections: among the projections onto five factors, only one is unable to estimate the quadratic model. Xu et al. (2004) were able to find orthogonal arrays, OA(27, $3^{13}$, 2), such that the quadratic model can be estimated in all of the five-factor projections.

## 5 Projection and Search Designs

The need to consider projections also arises in the study of search designs (Srivastava, 1975) that can be used to identify and estimate nonnegligible effects. Suppose that, in addition to a set of effects (say, the main effects of all factors) that must be estimated, there is another set of effects (say, all the two-factor interactions) that is known to the experimenter to contain at most a small number of nonnegligible effects, but the identity of these is not known. The objective of a search design is to find the nonnegligible effects and allow them to be estimated. This approach can be adapted for factor screening. Suppose that there are at most $p$ active factors among $f$ potentially active factors, where $p$ is known and is small, but it is unknown which of the factors are active. Here, as mentioned before, a factor is active if at least one factorial effect involving it is nonnegligible. Then the same kind of argument as that used by Srivastava (1975) can be applied to show that, in the noiseless case (no random error), a design of size $n$ can be used to identify the active factors and estimate all of their main effects and interactions if and only if

$$[\mathbf{1}_n \vdots \boldsymbol{X}(A_1, A_2)] \text{ is of full column rank,} \tag{1}$$

for all pairs $(A_1, A_2)$, where each of $A_1$ and $A_2$ is a subset of $p$ potentially active factors, $\mathbf{1}_n$ is the $n \times 1$ vector of 1s, and $\boldsymbol{X}(A_1, A_2)$ is the matrix whose columns

are contrasts that define the main effects of the factors in $A_1 \cup A_2$, the interactions of factors in $A_1$, and the interactions of factors in $A_2$. If, for example, it can be assumed that all interactions involving three or more factors are negligible, then in $\boldsymbol{X}(A_1, A_2)$ we need only to include two-factor interaction contrasts. For example, if $p = 3$, and $A_1$ is the subset of factors 1, 2, 3 from the design in Table 2 and $A_2$ is the subset of factors 3, 4, 5, then $\boldsymbol{X}(A_1, A_2)$ has columns $\boldsymbol{d}_1, \boldsymbol{d}_2, \boldsymbol{d}_3, \boldsymbol{d}_4, \boldsymbol{d}_5, \boldsymbol{d}_1 \odot \boldsymbol{d}_2, \boldsymbol{d}_1 \odot \boldsymbol{d}_3, \boldsymbol{d}_2 \odot \boldsymbol{d}_3, \boldsymbol{d}_3 \odot \boldsymbol{d}_4, \boldsymbol{d}_3 \odot \boldsymbol{d}_5, \boldsymbol{d}_4 \odot \boldsymbol{d}_5$.

When $p = 2$, $A_1 \cup A_2$ involves at most four factors. It was mentioned in the previous section that any OA($n$, $2^f$, 2) such that $n$ is not a multiple of 8, and any Plackett–Burman design of size $n \geq 12$, has the property that the main effects and two-factor interactions can be estimated in all four-factor projections. So, in these designs, (1) is satisfied and can, therefore, be used to identify up to two active factors and estimate their main effects and interaction. Similarly, the six-factor projection property of the 32-run Plackett–Burman design mentioned in Section 4 implies that the design can be used to identify up to $p = 3$ active factors and estimate their main effects and two-factor interactions, under the assumption that the three-factor and higher-order interactions are negligible. The reason why a design capable of searching for up to $p$ factors requires hidden projection properties for $2p$ factors is that, without the above full rank condition for $2p$ factors, identifiability issues would cause a pair of models each involving $p$ factors to be indistinguishable in certain situations.

Ghosh (1979), Ghosh and Avila (1985), and Müller (1993) considered the factor screening problem under the assumption that all the interactions are negligible. In this case, suppose there are $f$ two-level potentially active factors. For a design of size $n$, let $\boldsymbol{D}$ be the $n \times f$ matrix such that the $(i, j)$th entry is equal to 1 or $-1$ depending on whether, on the $i$th run, the $j$th factor is at the high or low level ($i = 1, \ldots, n$; $j = 1, \ldots, f$). Then the design can be used for identifying up to $p$ factors if and only if

$$\text{rank}[\mathbf{1} \vdots D_0] = 1 + 2p, \tag{2}$$

for every $n \times 2p$ submatrix $\boldsymbol{D}_0$ of $\boldsymbol{D}$.

If $[\mathbf{1} \vdots \boldsymbol{D}]$ is of full column rank, then clearly (2) is satisfied. Therefore, in this context, it is of interest to consider those designs with $f > n - 1$. These are called *supersaturated designs* which are discussed in Chapter 8.

## 6 Projection and Uniform Designs

Considerations of projections also arise in *uniform* or *space-filling designs*. Such designs are proposed in several settings, including computer experiments, numerical integration, robust regression, and drug discovery that call for the design points to be well spread over the entire design region. For discussions of drug discovery, see Chapter 4.

When the dimension is high, due to the curse of dimensionality, uniformity throughout the experimental region becomes difficult to achieve. Uniformity in

low-dimensional projections then provides a viable alternative. For example, in designs for computer experiments and numerical integration, Latin hypercube designs have been proposed as improvements on the random selection of design points (McKay et al., 1979). Such designs achieve better uniformity in one-dimensional projections: if $n$ points are to be selected from the unit cube $[0, 1]^f$ then, in all the univariate margins, exactly one point can be found in each of $n$ intervals of equal width. Extensions such as orthogonal array-based Latin hypercubes (Tang, 1993), randomized orthogonal arrays (Owen, 1992), and scrambled nets (Owen, 1997) achieve uniformity in low-dimensional (for example, binary and ternary) margins. Such designs are suitable when there are relatively few active factors.

Another approach to constructing uniform designs is based on minimizing a criterion of "discrepancy" which is a measure of the difference between the empirical distribution of the design points and the theoretical uniform distribution (Fang, 1980). In recent years, modifications have been proposed to take discrepancy in low-dimensional projections into account (Hickernell, 1998). Space-filling designs are also obtained through use of criteria that consider the coverage or spread of a design across a given set of candidate points. A projection approach based on these types of criteria is presented by Lam et al. (2002) for molecule selection in screening for drug discovery.

## 7 Discussion

In designs for factor screening, the number of factors is usually large, but the number of experimental runs is limited. The design points cannot provide a good coverage of the experimental region, nor can they support complicated models. On the other hand, often only a small number of factors and effects are expected to be important, and so it is sensible to use a design with good low-dimensional projections. This means that, when the design is restricted to the small subset of factors identified to be important, the design points have a good coverage of the reduced design space or are capable of entertaining more complicated models in the smaller set of factors. This is useful and important for the subsequent follow-up analyses and experiments. Knowledge about projection properties can also help the experimenter to incorporate prior knowledge into designs for factor-screening experiments; for example, it may be suspected that some factors are more likely to be important than others.

This chapter provides a survey of recent results on the projection properties of some commonly used designs, in particular, regular fractional factorial designs and nonregular designs such as the Plackett–Burman designs. The properties of the projection designs of a regular design can easily be determined from its defining relation, and so this chapter has concentrated more on nonregular designs whose complex alias structures lead to interesting hidden projection properties. The issue of how to identify active factors is discussed in Chapters 8 and 11. The reader is referred to Wu and Hamada (2000, Chapter 8), Cheng and Wu (2001), and the discussants' comments on the latter, for proposals and discussions of data

analysis, including critiques of, and alternatives to, the strategy of screening only main effects in the first stage of experimentation.

*Acknowledgment.* This work was supported by National Science Foundation Grant DMS-0071438.

## *References*

Box, G. E. P. and Bisgaard, S. (1993). What can you find out from 12 experimental runs? *Quality Engineering*, **5**, 663–668.

Box, G. E. P. and Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics*, **28**, 11–18.

Box, G. E. P. and Tyssedal, J. (1996). Projective properties of certain orthogonal arrays. *Biometrika*, **83**, 950–955.

Box, G. E. P. Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. John Wiley and Sons, New York.

Bulutoglu, D. A. and Cheng, C. S. (2003). Hidden projection properties of some nonregular fractional factorial designs and their applications. *Annals of Statistics*, **31**, 1012–1026.

Chen, H. (1998). Some projective properties of fractional factorial designs. *Statistics and Probability Letters*, **40**, 185–188.

Cheng, C. S. (1995). Some projection properties of orthogonal arrays. *Annals of Statistics*, **23**, 1223–1233.

Cheng, C. S. (1998a). Hidden projection properties of orthogonal arrays with strength three. *Biometrika*, **85**, 491–495.

Cheng, C. S. (1998b). Projectivity and resolving power. *Journal of Combinatorics, Information and System Sciences*, **23**, 47–58.

Cheng, S. W. and Wu, C. F. J. (2001). Factor screening and response surface exploration (with discussions). *Statistica Sinica*, **11**, 553–604.

Constantine, G. M. (1987). *Combinatorial Theory and Statistical Design*. John Wiley and Sons, New York.

Dey, A. and Mukerjee, R. (1999). *Fractional Factorial Plans*. John Wiley and Sons, New York.

Fang, K. T. (1980). The uniform design: Application of number–theoretic methods in experimental design. *Acta Mathematicae Applicatae Sinica*, **3**, 363–372.

Ghosh, S. (1979). On single and multistage factor screening procedures. *Journal of Combinatorics, Information and System Sciences*, **4**, 275–284.

Ghosh, S. and Avila, D. (1985). Some new factor screening designs using the search linear model. *Journal of Statistical Planning and Inference*, **11**, 259–266.

Hamada, M. and Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, **24**, 130–137.

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999). *Orthogonal Arrays: Theory and Applications*. Springer–Verlag, New York.

Hickernell, F. J. (1998). A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, **67**, 299–322.

Lam, R. L. H., Welch, W. J., and Young, S. S. (2002). Uniform coverage designs for molecule selection. *Technometrics*, **44**, 99–109.

Lin, D. K. J. and Draper, N. R. (1991). Projection properties of Plackett and Burman designs. Technical Report 885, Department of Statistics, University of Wisconsin.

Lin, D. K. J. and Draper, N. R. (1992). Projection properties of Plackett and Burman designs. *Technometrics*, **34**, 423–428.

Lin, D. K. J. and Draper, N. R. (1993). Generating alias relationships for two-level Plackett and Burman designs. *Computational Statistics and Data Analysis*, **15**, 147–157.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239–245.

Müller, M. (1993). Supersaturated designs for one or two effective factors. *Journal of Statistical Planning and Inference*, **37**, 237–244.

Owen, A. B. (1992). Orthogonal arrays for computer experiments, integration and visualization. *Statistics Sinica*, **2**, 439–452.

Owen, A. B. (1997). Scrambled net variance for integrals of smooth functions. *Annals of Statistics*, **25**, 1541–1562.

Paley, R. E. A. C. (1933). On orthogonal matrices. *Journal of Mathematical Physics*, **12**, 311–320.

Plackett, R. L. and Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, **33**, 305–325.

Rao, C. R. (1947). Factorial experiments derivable from combinatorial arrangements of arrays. *Journal of Royal Statistical Society, Supplement* **9**, 128–139.

Sloane, N. J. A. (1993). Covering arrays and intersecting codes. *Journal of Combinatorial Designs*, **1**, 51–63.

Srivastava, J. N. (1975). Designs for searching non-negligible effects. In *A Survey of Statistical Design and Linear Models*. Editor: J. N. Srivastava, pages 507–519. Elsevier, Amsterdam.

Tang, B. (1993). Orthogonal array-based Latin hypercubes. *Journal of the American Statistical Association*, **88**, 1392–1397.

Wang, J. C. and Wu, C. F. J. (1995). A hidden projection property of Plackett–Burman and related designs. *Statistica Sinica*, **5**, 235–250.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992). Screening, predicting, and computer experiments. *Technometrics*, **34**, 15–25.

Wu, C. F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley and Sons, New York.

Xu, H., Cheng, S. W., and Wu, C. F. J. (2004). Optimal projective three-level designs for factor screening and interaction detection. *Technometrics*, **46**, 280–292.

# 8
# Factor Screening via Supersaturated Designs

STEVEN G. GILMOUR

Supersaturated designs are fractional factorial designs that have too few runs to allow the estimation of the main effects of all the factors in the experiment. There has been a great deal of interest in the development of these designs for factor screening in recent years. A review of this work is presented, including criteria for design selection, in particular the popular $E(s^2)$ criterion, and methods for constructing supersaturated designs, both combinatorial and computational. Various methods, both classical and partially Bayesian, have been suggested for the analysis of data from supersaturated designs and these are critically reviewed and illustrated. Recommendations are made about the use of supersaturated designs in practice and suggestions for future research are given.

## 1 Supersaturated Designs

All factor screening designs are intended for situations in which there are too many factors to study in detail. If the number of factors is very large and/or each experimental run is very expensive, then it may be impractical to use even the Resolution III two-level designs of Chapter 1, which allow all main effects to be estimated. In such cases, it might be useful to run experiments with fewer runs than there are factors to try to identify a small number of factors that appear to have dominant effects.

A *supersaturated* design is a design for which there are fewer runs than effects to be estimated in a proposed model. Usually the term "supersaturated" is used more specifically for a design for which there are insufficient degrees of freedom for estimating the main effects only model, that is, for designs with $n$ runs where estimating the main effects would require more than $n - 1$ degrees of freedom. This chapter discusses the use of screening designs of this more restricted type. In the early literature on the subject, and often today, the term was used still more specifically for a design with $n$ runs in which there are more than $n - 1$ *two-level* factors. The discussion here concentrates on these two-level designs and their use in factor screening. For recent work on factors with more than two levels, see Lu et al. (2003) and the references contained therein.

Supersaturated designs have their roots in *random balance experimentation*, which was briefly popular in industry in the 1950s, until the discussion of the papers

by Satterthwaite (1959) and Budne (1959). In these experiments, the combinations of the factor levels are chosen at random, subject to having equal numbers of runs at each level of each factor, and they can include more factors than there are runs. Box (1959) suggested that the latter idea was worth pursuing in the context of designed experiments. However, the idea of random balance itself was totally refuted as a useful way of running experiments and has rarely been seen since. Booth and Cox (1962) presented the first supersaturated designs, but no more work on the subject was published for more than 30 years.

The papers by Lin (1993) and Wu (1993) sparked a renewed interest in the subject. Since then there has been a large and increasing number of papers published in the statistical literature, mostly on methods of constructing supersaturated designs. It is less clear how much they are being used in practice. There appear to be no published case studies featuring the use of supersaturated designs, although the most likely area for their application is in early discovery experiments which are unlikely to be sent for publication. In my own experience, industrial statisticians are reluctant to recommend supersaturated designs because there are no successful case studies in the literature, but the difficulties in interpreting the data might also be a deterrent.

This chapter reviews the recent work on supersaturated designs in factor screening, concentrating on methods of obtaining designs and analyzing the data that are most likely to be useful in practice. It also attempts to assess how much we know about the usefulness of supersaturated designs and suggests areas where more research is needed. Section 2 reviews methods of constructing designs, and Section 3 discusses the methods of analysis that have been recommended. In Section 4, some recommendations for future research are made, including comparison of supersaturated designs with alternatives, exploratory data analysis, and Bayesian modeling. A brief discussion is given in Section 5.

## 2 $E(s^2)$-Optimal Designs

We assume that each factor has two levels, coded $+1$ and $-1$, often written as $+$ and $-$. As in almost all of the literature, we assume that each factor is observed at each level an equal number of times, although Allen and Bernshteyn (2003) recently relaxed this assumption.

### *2.1 Criteria of Optimality*

Consider the "main effects only" model,

$$Y = \beta_0 + \sum_{j=1}^{f} \beta_j x_j + \epsilon, \tag{1}$$

where $Y$ is a response variable, $\beta_0, \beta_1, \ldots, \beta_f$ are unknown parameters, $x_1, \ldots, x_f$ are the coded levels of the $f$ factors, $\epsilon$ is an error term with $E(\epsilon) = 0$ and

TABLE 1. Design for 10 factors in 6 runs.

| Factors | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| + | + | − | + | + | + | − | − | − | + |
| + | + | + | − | − | − | + | − | + | + |
| − | − | + | − | + | + | − | + | + | + |
| − | + | − | + | + | − | + | + | + | − |
| + | − | + | + | − | + | + | + | − | − |
| − | − | − | − | − | − | − | − | − | − |

$V(\epsilon) = \sigma^2$, and error terms are independent. We also write this model in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2}$$

In a supersaturated design, even for this main effects only model, the matrix $\boldsymbol{X}'\boldsymbol{X}$ is singular, where $'$ denotes transpose, and so no unique least squares estimates of the parameters $\boldsymbol{\beta}$ can be obtained. Consider, for example, the small supersaturated design shown in Table 1. This has

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6 & 2 & 2 & 2 & -2 & 2 & 2 & -2 & -2 & 2 \\ 0 & 2 & 6 & -2 & 2 & 2 & -2 & 2 & -2 & 2 & 2 \\ 0 & 2 & -2 & 6 & -2 & -2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 2 & 2 & -2 & 6 & 2 & 2 & 2 & 2 & -2 & -2 \\ 0 & -2 & 2 & -2 & 2 & 6 & 2 & -2 & 2 & 2 & 2 \\ 0 & 2 & -2 & 2 & 2 & 2 & 6 & -2 & 2 & -2 & 2 \\ 0 & 2 & 2 & 2 & 2 & -2 & -2 & 6 & 2 & 2 & -2 \\ 0 & -2 & -2 & 2 & 2 & 2 & 2 & 2 & 6 & 2 & -2 \\ 0 & -2 & 2 & 2 & -2 & 2 & -2 & 2 & 2 & 6 & 2 \\ 0 & 2 & 2 & 2 & -2 & 2 & 2 & -2 & -2 & 2 & 6 \end{bmatrix}. \tag{3}$$

The diagonal elements of this matrix are fixed by the number of runs. If it were possible to reduce the off-diagonal elements in absolute value, the matrix could be made nonsingular. Best of all would be if the off-diagonal elements were all zero, in which case all main effects would be estimated independently.

It is, of course, impossible to get an $\boldsymbol{X}'\boldsymbol{X}$ matrix for 10 factors in 6 runs with rank greater than 6. However, these considerations suggest that a good design will be one that makes the off-diagonal elements as small as possible (in absolute value). Letting the $(i, j)$th element of $\boldsymbol{X}'\boldsymbol{X}$ be $s_{ij}$, Booth and Cox (1962) suggested two criteria based on the sizes of the $s_{ij}$. The first criterion they used was to choose a design with minimum $\max_{i \neq j} |s_{ij}|$, and among all such designs to choose one with the fewest $s_{ij}$ that achieves this maximum.

The second suggestion of Booth and Cox was to choose a design that minimizes

$$E(s^2) = \frac{2}{f(f-1)} \sum_{i<j} s_{ij}^2. \tag{4}$$

This has become the most commonly used criterion in the literature on supersaturated designs. Sometimes these two criteria are combined, for example, by choosing a supersaturated design that minimizes $E(s^2)$ subject to some upper bound on $\max_{i \neq j} |s_{ij}|$. Cheng and Tang (2001) gave upper bounds on the number of factors that could be included in a supersaturated design subject to $\max_{i \neq j} |s_{ij}| \leq c$, where $c$ is a constant.

Booth and Cox (1962) gave two other interpretations of $E(s^2)$. They showed that if there are only $p$ important factors and their main effects are so large that they can be easily identified, then the average variance of their estimated main effects is approximately

$$\frac{\sigma^2}{n} \left\{ 1 + \frac{(p-1)E(s^2)}{n^2} \right\}. \tag{5}$$

Thus, an $E(s^2)$-optimal design also minimizes this quantity. As $p$ becomes larger, this approximation becomes poorer and the assumption that the large effects can be identified becomes less plausible, so it is most relevant for $p = 2$ and perhaps $p = 3$. Wu (1993) showed that $E(s^2)$-optimal designs also maximize the average $D$-efficiency over all models with just two main effects.

The second interpretation of $E(s^2)$ arises from considering the estimation of the main effect of a single factor $X_j$, for example if only one factor appears to have a very large effect. The simple linear regression estimate of $\beta_j$ from the model $Y_i = \beta_0 + \beta_j x_{ji} + \epsilon_i$

$$\hat{\beta}_j = \sum_{i=1}^{n} x_{ji} y_i / n, \text{ with } V(\hat{\beta}_j) = \sigma^2 / n,$$

and is based on the assumption that all other factors will have zero effects. If, in fact, all other factors have effects of magnitude $2\delta$, with their directions being chosen at random, then the true variance of the single estimated main effect is not $\sigma^2/n$, but

$$\frac{\sigma^2}{n} + \frac{(f-1)}{4n^2} \delta^2 E(s^2). \tag{6}$$

Cheng et al. (2002) showed that supersaturated designs with a property called "minimum $G_2$-aberration" are $E(s^2)$-optimal and suggested that $G_2$-aberration might be a useful criterion for supersaturated designs. Liu and Hickernall (2002) showed that $E(s^2)$ is similar, but not identical, to a form of *discrepancy*, that is, a measure of how far the points of the design are from being uniformly distributed in the factor space. They also showed that, under certain conditions, the most uniform designs are $E(s^2)$-optimal. It is unknown how the concept of discrepancy is related to the statistical properties of the designs.

A different criterion was used by Allen and Bernshteyn (2003) to construct supersaturated designs. They used prior probabilities of factors being active (having

nonnegligible effects) and inactive (having negligible effects) and then searched for designs that maximize the probability of correctly selecting the active factors. In all of the examples they studied, they found that designs which optimize this criterion are also $E(s^2)$-optimal, but that the converse is not true. This suggests that they could restrict their search for designs to the class of $E(s^2)$-optimal designs.

Other criteria have been suggested, but rarely used, for constructing supersaturated designs, although they are sometimes used for comparing different $E(s^2)$-optimal designs. One of these is to minimize the average $D$- or $A$-efficiency over all submodels with $p$ factors, $2 \leq p < f$ (Wu, 1993; Liu and Dean, 2004). Deng et al. (1996) suggested using the multiple regression of a column in the design on $p-1$ other columns. The regression sum of squares then gives a measure of the nonorthogonality of this column to the others. The average of this regression sum of squares over all sets of columns can be used as a criterion for comparing designs, although the computation of this criterion is a major task in itself. Deng et al. (1999) defined the resolution-rank of a supersaturated design as the maximum $p$ such that any $p$ columns of the design are linearly independent. They suggested maximizing the resolution-rank, although again the computation of this criterion is prohibitive for large $f$. Holcomb and Carlyle (2002) suggested using the ratio the largest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$ and the smallest nonzero eigenvalue and stated that this was related to $A$-efficiency. None of these criteria have been studied further.

## 2.2 *Methods for Constructing $E(s^2)$-Optimal Designs*

Several different methods of constructing supersaturated designs have been suggested, most based on Hadamard matrices, incomplete block designs, or computer search routines. In the earlier papers, when a new method was suggested, it was shown to give designs that are better than those obtained by previous methods with regard to the $E(s^2)$ criterion. Some of these designs have later been shown to be $E(s^2)$-optimal, whereas others have been shown to be $E(s^2)$-suboptimal. In this section, attention is paid only to methods that are known to lead to $E(s^2)$-optimal designs. Ways of constructing designs when no $E(s^2)$-optimal design is known are discussed in the next section.

In order to know whether a given design is optimal, it is helpful to have lower bounds on $E(s^2)$. Increasingly tight, or more widely applicable, bounds have been given by Nguyen (1996), Tang and Wu (1997), Liu and Zhang (2000a,b), Butler et al. (2001), and Bulutoglu and Cheng (2004). The bounds of Bulutoglu and Cheng cover all cases with $n$ even and with each factor having two levels with $n/2$ runs at each level, that is, all of the cases we are considering here. These results allow us to identify many $E(s^2)$-optimal supersaturated designs.

### 2.2.1 Methods Using Hadamard Matrices

A $t \times t$ matrix $\boldsymbol{H}$ with elements $\pm 1$ is called a Hadamard matrix if $\boldsymbol{H}'\boldsymbol{H} = t\boldsymbol{I}$. Hadamard matrices are considered to be equivalent if they can be obtained from

TABLE 2. Plackett–Burman design for 11 factors in 12 runs.

| Factors | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| + | + | − | + | + | + | − | − | − | + | − |
| + | − | + | + | + | − | − | − | + | − | + |
| − | + | + | + | − | − | − | + | − | + | + |
| + | + | + | − | − | − | + | − | + | + | − |
| + | + | − | − | − | + | − | + | + | − | + |
| + | − | − | − | + | − | + | + | − | + | + |
| − | − | − | + | − | + | + | − | + | + | + |
| − | − | + | − | + | + | − | + | + | + | − |
| − | + | − | + | + | − | + | + | + | − | − |
| + | − | + | + | − | + | + | + | − | − | − |
| − | + | + | − | + | + | + | − | − | − | + |
| − | − | − | − | − | − | − | − | − | − | − |

each other by elementary row operations and so they can always be written with the first column consisting entirely of 1. Nguyen (1996) showed that the following method of construction, due to Lin (1993), gives $E(s^2)$-optimal supersaturated designs. Take a $2n \times 2n$ Hadamard matrix, $\boldsymbol{H}$, and write it as

$$\boldsymbol{H} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \boldsymbol{H}_1 \\ \mathbf{1} & -\mathbf{1} & \boldsymbol{H}_2 \end{bmatrix}, \tag{7}$$

where $\mathbf{1}$ is an $n \times 1$ vector with every element 1. Then $\boldsymbol{H}_1$ is an $E(s^2)$-optimal supersaturated design for $2(n-1)$ factors in $n$ runs if it contains no identical columns.

This method can be illustrated by taking the Plackett–Burman design for 11 factors in 12 runs, which is based on a $12 \times 12$ Hadamard matrix and is shown in Table 2. Selecting the rows corresponding to + in Factor 11 and dropping this factor gives the supersaturated design for 10 factors in 6 runs shown in Table 1.

Deng et al. (1994) suggested using quarter, eighth, or smaller fractions of Hadamard matrices and Cheng (1997) showed that these are also $E(s^2)$-optimal. Tang and Wu (1997) showed that $E(s^2)$-optimal designs can be constructed by joining Hadamard matrices. Suppose that $m$ Hadamard matrices, $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_m$, each of size $n \times n$, exist with no columns in common. Then, writing $\boldsymbol{H}_i$ as $[\mathbf{1} \quad \boldsymbol{H}_i^*]$, the array $[\boldsymbol{H}_1^* \cdots \boldsymbol{H}_m^*]$ gives an $E(s^2)$-optimal supersaturated design for $m(n-1)$ factors in $n$ runs. For example, the design shown in Table 3 combines that in Table 2 with another Hadamard matrix obtained by permuting its rows and has no repeated columns. Hence it is an $E(s^2)$-optimal supersaturated design for 22 factors in 12 runs.

Wu (1993) suggested another method of construction, namely, adding interaction columns to a saturated design obtained from a Hadamard matrix with the first column deleted. For example, adding the pairwise interactions of Factor 1 with every other factor to the design in Table 2, we get the supersaturated design for 21 factors in 12 runs given in Table 4. Bulutoglu and Cheng (2003) showed that

TABLE 3. Design for 22 factors in 12 runs.

| Factors | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| + | + | − | + | + | + | − | − | − | + | − | + | − | − | − | + | − | + | + | − | + | + |
| + | − | + | + | + | − | − | − | + | − | + | − | − | − | + | − | + | + | − | + | + | + |
| − | + | + | + | − | − | − | + | − | + | + | − | − | + | − | + | + | − | + | + | + | − |
| + | + | + | − | − | − | + | − | + | + | − | − | + | − | + | + | − | + | + | + | − | − |
| + | + | − | − | − | + | − | + | + | − | + | + | − | + | + | + | − | − | − | + | − | + |
| + | − | − | − | + | − | + | + | − | + | + | − | + | + | + | − | − | − | + | − | + | + |
| − | − | − | + | − | + | + | − | + | + | + | + | + | + | − | − | − | + | − | + | + | − |
| − | − | + | − | + | + | − | + | + | + | − | + | + | − | − | − | + | − | + | + | − | + |
| − | + | − | + | + | − | + | + | + | − | − | + | − | + | + | − | + | + | + | − | − | − |
| + | − | + | + | − | + | + | + | − | − | − | − | + | + | − | + | + | + | − | − | − | + |
| − | + | + | − | + | + | + | − | − | − | + | + | + | − | + | + | + | − | − | − | + | − |
| − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − |

this procedure does not, in general, produce $E(s^2)$-optimal supersaturated designs, but that it does in the following case.

- All interactions are used, along with the original design;
- All interactions are used, without the original design;
- All interactions with one factor are used, along with the original design.

Thus the design shown in Table 4 is $E(s^2)$-optimal.

### 2.2.2 Method Using Incomplete Block Designs

Nguyen (1996) showed that $E(s^2)$-optimal supersaturated designs could be obtained from balanced incomplete block designs for $n-1$ treatments in blocks of size $n/2-1$, provided that there are no repeated blocks. Each block generates a

TABLE 4. Design for 21 factors in 12 runs.

| Factors | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | |
| + | + | − | + | + | + | − | − | − | + | − | + | − | + | + | + | − | − | − | + | − | |
| + | − | + | + | + | − | − | − | + | − | + | − | + | + | + | − | − | − | + | − | + | |
| − | + | + | + | − | − | − | + | − | + | + | − | − | − | + | + | + | − | + | − | − | |
| + | + | + | − | − | − | + | − | + | + | − | + | + | − | − | − | + | − | + | + | − | |
| + | + | − | − | − | + | − | + | + | − | + | + | − | − | − | + | − | + | + | − | + | |
| + | − | − | − | + | − | + | + | − | + | + | − | − | − | + | − | + | + | − | + | + | |
| − | − | − | + | − | + | + | − | + | + | + | + | + | − | + | − | − | + | − | − | − | |
| − | − | + | − | + | + | − | + | + | + | − | + | − | + | − | − | + | − | − | − | + | |
| − | + | − | + | + | − | + | + | + | − | − | − | + | − | − | + | − | − | − | + | + | |
| + | − | + | + | − | + | + | + | − | − | − | − | + | + | − | + | + | + | − | − | − | |
| − | + | + | − | + | + | + | − | − | − | + | − | − | + | − | − | − | + | + | + | − | |
| − | − | − | − | − | − | − | − | − | − | − | + | + | + | + | + | + | + | + | + | + | |

TABLE 5. Design for 14 factors in 8 runs.

| Factors | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| − | + | − | − | − | + | + | − | − | − | + | − | + | + |
| + | − | + | − | − | − | + | + | − | − | − | + | − | + |
| + | + | − | + | − | − | − | + | + | − | − | − | + | − |
| − | + | + | − | + | − | − | − | + | + | − | − | − | + |
| − | − | + | + | − | + | − | + | − | + | + | − | − | − |
| − | − | − | + | + | − | + | − | + | − | + | + | − | − |
| + | − | − | − | + | + | − | − | − | + | − | + | + | − |
| + | + | + | + | + | + | + | + | + | + | + | + | + | + |

column of the supersaturated design with a + in row $i$ if treatment $i$ is in the block. A final row of + is added to complete the supersaturated design. For example, one possible balanced incomplete block design for 7 treatments in 14 blocks of size 3 has the following blocks: (2,3,7), (1,3,4), (2,4,5), (3,5,6), (4,6,7), (1,5,7), (1,2,6), (2,3,5), (3,4,6), (4,5,7), (1,5,6), (2,6,7), (1,3,7), and (1,2,4). This produces the supersaturated design for 14 factors in 8 runs shown in Table 5. Because every Hadamard matrix is equivalent to a balanced incomplete block design, but not every balanced incomplete block design is equivalent to a Hadamard matrix, Bulutoglu and Cheng (2004) pointed out that construction of supersaturated designs using balanced incomplete block designs is more flexible. The only problem is in ensuring that there are no repeated blocks. Bulutoglu and Cheng gave a number of results useful in constructing balanced incomplete block designs with no repeated blocks.

### 2.2.3 Methods Using Known $E(s^2)$-Optimal Designs

The above construction methods give designs for many values of $n$, the number of runs, and $f$, the number of factors, but there are still many values of $n$ and $f$ for which they do not provide designs. Fortunately, methods have been developed for constructing new $E(s^2)$-optimal designs based on known $E(s^2)$-optimal designs.

Nguyen (1996) showed that deleting any single column from an $E(s^2)$-optimal supersaturated design obtained from a balanced incomplete block design results in an $E(s^2)$-optimal supersaturated design. Cheng (1997) showed how to obtain $E(s^2)$-optimal supersaturated designs for $m(n-1)\pm 1$ and $m(n-1)\pm 2$ factors in $n$ runs when an $E(s^2)$-optimal design is known for $m(n-1)$ factors in $n$ runs for any positive integer $m$. The addition of any one column that does not already appear, or deleting any one column from the design, gives an $E(s^2)$-optimal design. Hence deleting any column from the design in Table 3 gives an alternative $E(s^2)$-optimal design for 21 factors in 12 runs to that shown in Table 4.

The cases of $f = m(n-1)\pm 2$ factors require more care. If $n$ is a multiple of 4, then we can add or delete any pair of *orthogonal* columns to obtain an

$E(s^2)$-optimal design; that is, in the two columns added or deleted each combination $(-1, -1), (-1, +1), (+1, -1), (+1, +1)$, must appear $n/4$ times. If $n$ is not a multiple of 4, then the two columns added or deleted must be *nearly* orthogonal in the sense that each of $(-1, -1)$ and $(1, 1)$ appears $(n + 2)/4$ times and each of $(-1, 1)$ and $(1, -1)$ appears $(n - 2)/4$ times.

Butler et al. (2001) gave several methods for obtaining $E(s^2)$-optimal supersaturated designs from smaller $E(s^2)$-optimal supersaturated designs. First, if

1. $\boldsymbol{D}_0$ is an $E(s^2)$-optimal supersaturated design for $f_0$ factors in $n$ runs,
2. $\boldsymbol{D}_1$ is an $E(s^2)$-optimal supersaturated design for $f_1 = m(n - 1)$ factors in $n$ runs, where $m$ must be even if $n \equiv 2(\text{mod } 4)$, and
3. $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ have no columns in common,

then $[\boldsymbol{D}_0 \ \ \boldsymbol{D}_1]$ is an $E(s^2)$-optimal supersaturated design for $f_0 + f_1$ factors in $n$ runs. The method of Tang and Wu (1997), described in Section 2.2, uses a special case of this result.

Secondly, if

1. $\boldsymbol{D}_0$ is an $E(s^2)$-optimal supersaturated design for $f_0$ factors in $n/2$ runs, where
$$f_0 = m\left(\frac{n}{2} - 1\right) + \text{sign}(r)\left\{|r| + 4\text{int}\left(\frac{|r|}{8}\right) - 4\text{int}\left(\frac{|r|}{4}\right)\right\}, \tag{8}$$
where $\text{sign}(\cdot)$ and $\text{int}(\cdot)$ denote, respectively, the sign of the argument and the integer part of the argument, where $r$ is an integer;
2. $\boldsymbol{D}_1$ is a design for $f_1$ factors in $n/2$ runs with orthogonal rows, where
$$f_1 = \frac{mn}{2} + 4\text{sign}(r)\text{int}\left(\frac{|r| + 4)}{8}\right); \tag{9}$$
3. $f_0 + f_1 = m(n - 1) + r$, where $|r| < n/2$; and
4. either
   - $n$ is a multiple of 8 and $|r| \not\equiv 3(\text{mod } 8)$; or
   - $n$ is a multiple of 4 but not 8, $m$ is even, $|r| \not\equiv 2(\text{mod } 4)$, and $|r| \not\equiv 3(\text{mod } 8)$,

then
$$\begin{bmatrix} \boldsymbol{D}_0 & \boldsymbol{D}_1 \\ \boldsymbol{D}_0 & -\boldsymbol{D}_1 \end{bmatrix} \tag{10}$$
is an $E(s^2)$-optimal design for $f_0 + f_1$ factors in $n$ runs.

Thirdly, any design obtained by deleting any column from a design obtained by either of the two methods just described, or adding any column distinct from those already in the design, is an $E(s^2)$-optimal design. This extends the result of Cheng (1997) to more general numbers of factors.

Finally, if

1. $\boldsymbol{D}_0$ is an $E(s^2)$-optimal supersaturated design for $f_0$ factors in $n/2^k$ runs where
$$f_0 = \left(\frac{mn}{2^k} - 1\right) + \text{sign}(r)\left\{|r| + 4\text{int}\left(\frac{|r|}{2^{k+2}}\right) - 4\text{int}\left(\frac{|r|}{4}\right)\right\}; \tag{11}$$

2. $\boldsymbol{D}_i, i = 1, \ldots, k$, is a design for $f_i$ factors in $n/2^i$ runs with orthogonal rows, where

$$f_i = \frac{mn}{2^i} + 4\text{sign}(r)\text{int}\left(\frac{|r|}{2^{i+2}} + \frac{1}{2}\right); \tag{12}$$

and

3. either
   - $n$ is a multiple of $2^{k+2}$ and $|r| \not\equiv 3(\text{mod } 4)$; or
   - $n$ is a multiple of $2^{k+1}$ but not of $2^{k+2}$, $m$ is even, and $|r| \equiv 0$ or $1(\text{mod } 4)$,

then

$$[a_1 \otimes \boldsymbol{D}_1 \ldots a_k \otimes \boldsymbol{D}_k \quad 1_{2^k} \otimes \boldsymbol{D}_0] \tag{13}$$

is an $E(s^2)$-optimal supersaturated design for $f_0 + f_1 + \cdots + f_k$ factors in $n$ runs, where $a_i$ has length $2^i$ and has odd elements equal to 1 and even elements equal to $-1$.

### 2.2.4 Computer Construction

Several $E(s^2)$-optimal designs are known that cannot be obtained from any of the methods described above. Rather, they were constructed using computer programs. These programs are based on algorithms that are not guaranteed to find the best design, but the designs found can be compared with the corresponding lower bounds for $E(s^2)$ and sometimes they will achieve these bounds.

Perhaps surprisingly, the earliest supersaturated designs, those of Booth and Cox (1962), were obtained by computer-aided construction. However, their algorithm has been shown to produce suboptimal designs in many situations. Lin (1995) was the first to use a modern exchange algorithm, but the greatest advance was the algorithm of Nguyen (1996). He realized that it was only necessary to swap $+$ and $-$ within columns of a design, thus allowing an *interchange* algorithm to be used, which is much more efficient than the exchange algorithm. A similar algorithm was suggested by Li and Wu (1997).

Nguyen's algorithm, known as NOA, is implemented within the Gendex package; see http://www.designcomputing.net/gendex for further details. Bulutoglu and Cheng (2004) used this program to find several new $E(s^2)$-optimal designs. Cela et al. (2000) instead used a genetic algorithm to find $E(s^2)$-optimal designs, but this does not appear to have any advantage over the NOA algorithm.

## *2.3 Other Designs*

If no $E(s^2)$-optimal design is known to exist for a particular run size and number of factors, then it is sensible to try to find a nearly $E(s^2)$-optimal design. A reasonable approach is to use an algorithm such as NOA, because this might find a design that can be shown to reach the $E(s^2)$ lower bound but if it does not, the design it finds

might actually be $E(s^2)$-optimal and should certainly be nearly $E(s^2)$-optimal. However, this approach might fail to find a good design in a reasonable amount of computing time, when very large designs are needed.

Various other approaches to constructing nearly $E(s^2)$-optimal designs have been suggested. One is the construction of Wu (1993), described in Section 2.2, based on adding interaction columns to a Hadamard matrix which, when it does not produce $E(s^2)$-optimal designs, produces designs with reasonably high $E(s^2)$-efficiencies.

The correspondence between incomplete block designs and supersaturated designs applies to unbalanced, as well as balanced, incomplete block designs. For this reason, Eskridge et al. (2004) suggested using cyclic regular graph designs. They showed that these designs are at least 94.9% $E(s^2)$-efficient for $f = m(n-1)$ and $n \geq 10$ and showed how they can be used to construct supersaturated designs for up to 12,190 factors in 24 runs. Liu and Dean (2004) used a more general type of cyclic design, known as $k$-circulant supersaturated designs, to find all designs based on balanced incomplete block designs, all those of Eskridge et al. (2004), and several new designs. All the designs they presented are at least 97.8% $E(s^2)$-efficient.

Although several authors have used secondary criteria, especially minimizing $\max_{i \neq j} |s_{ij}|$, to distinguish between different $E(s^2)$-optimal designs, or to illustrate the properties of the designs, few have used other criteria as the principal aim of construction. One exception is Yamada and Lin (1997), who considered the situation where a subset of the factors is identified, a priori, to be more likely to have nonnegligible effects. They then built designs to minimize $E(s^2)$ subject to correlations between pairs of factors within the identified group being less than a given bound. Although appropriate for this purpose, the overall $E(s^2)$ of these designs is much higher than the lower bounds.

## 3 Data Analysis

Standard methods for analyzing data from fractional factorial designs cannot be used with data from supersaturated designs, because the least squares estimates are not unique and, given any reasonable assumptions, there is no way to estimate all the main effects simultaneously.

Several methods of analysis have been suggested in the literature and are discussed in the context of data from half of an experiment reported by Williams (1968) and analyzed by several authors. Twenty-three factors were varied in 28 runs and one continuous response was observed. The half-fraction analyzed by Lin (1993) is shown in Table 6, which incorporates the corrections noted by Box and Draper (1987) and Abraham et al. (1999).

Most methods of analysis assume that the objective is to identify a few active factors, those with nonnegligible main effects, to separate them from the inactive factors, those with negligible main effects.

TABLE 6. Design and data for half-replicate of Williams' experiment.

| Factors | | | | | | | | | | | | | | | | | | | | | | | Response |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | $y$ |
| + | + | + | − | − | − | + | + | + | + | + | − | + | − | − | + | − | − | + | − | − | − | + | 133 |
| + | − | − | − | − | − | + | + | + | − | − | − | + | + | + | − | + | − | − | + | + | − | − | 62 |
| + | + | − | + | + | − | − | − | − | + | − | + | + | + | + | + | − | − | − | − | + | + | − | 45 |
| + | + | − | + | − | + | − | − | − | + | + | − | + | − | + | − | + | + | + | − | − | − | − | 52 |
| − | − | + | + | + | + | − | + | + | − | − | − | + | − | + | + | − | − | + | − | + | + | + | 56 |
| − | − | + | + | + | + | + | − | + | + | + | − | − | + | + | + | + | + | + | + | + | − | − | 47 |
| − | − | − | − | + | − | − | + | − | + | − | + | + | + | − | + | + | + | + | + | − | − | + | 88 |
| − | + | + | − | − | + | − | + | − | + | − | − | − | − | − | − | − | + | − | + | + | + | − | 193 |
| − | − | − | − | − | + | + | − | − | − | + | + | − | − | + | + | + | − | − | − | − | + | + | 32 |
| + | + | + | + | − | + | + | + | − | − | − | + | − | + | + | + | − | + | − | + | − | − | + | 53 |
| − | + | − | + | + | − | − | + | + | − | + | − | − | + | − | − | + | + | − | − | − | + | + | 276 |
| + | − | − | − | + | + | + | − | + | + | + | + | + | − | − | − | − | + | − | + | + | + | + | 145 |
| + | + | + | + | + | − | + | − | + | − | − | + | − | − | − | − | + | − | + | + | − | + | − | 130 |
| − | − | + | − | − | − | − | − | − | − | + | + | − | + | − | − | − | − | + | − | + | − | − | 127 |

## 3.1 *Least Squares Estimation Methods*

Most often data analysis techniques are borrowed from regression analysis. Satterthwaite (1959) suggested a graphical method that is essentially equivalent to producing the least squares estimates from each simple linear regression. For example, in the data in Table 6, the effect of factor 1 is estimated by fitting $Y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$ and so on, giving the results in Table 7. Chen and Lin (1998) showed that, if there is a single active factor with true magnitude greater than $\sigma$ and if all other factors have exactly zero effect, then this procedure gives a high probability of the active factor having the largest estimated main effect. Lin (1995) suggested plotting these simple linear regression estimates on normal probability paper, although the lack of orthogonality makes the interpretation of such a plot difficult. Kelly and Voelkel (2000) showed that the probabilities of type-II errors resulting from this method are very high and recommended instead that all subsets selection be used, which involves fitting all estimable submodels of the main effects model.

Holcomb et al. (2003) studied the method of using the simple linear regression estimates in more detail. They described it as a contrast-based method obtained by using $\boldsymbol{X}'\boldsymbol{y}$ from the full model, but these are the simple linear regression estimates multiplied by $n$. They tried several procedures for separating the active from the

TABLE 7. Estimates of main effects obtained from the data of Table 6 using simple linear regressions

| Factors | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Effect | −14.2 | 23.2 | 2.8 | −8.6 | 9.6 | −20.2 | −16.8 | 20.2 | 18.5 | −2.4 | 13.2 | −14.2 |

| Factor | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Effect | −19.8 | −3.1 | −53.2 | −37.9 | −4.6 | 19.2 | −12.4 | −0.2 | −6.4 | 22.5 | 9.1 |

inactive factors based on these estimates. In a large simulation study they found that resampling methods to control the type-II error rate worked best and better than stepwise selection. However, they also concluded that none of the methods worked very well.

Lin (1993) suggested using stepwise variable selection and Wu (1993) suggested forward selection or all (estimable) subsets selection. Lin (1993) gave an illustrative analysis by stepwise selection of the data in Table 6. He found that this identified factors 15, 12, 19, 4, and 10 as the active factors, when their main effects are entered into the model in this order. Wang (1995) analyzed the other half of the Williams experiment and identified only one of the five factors that Lin had identified as being nonnegligible, namely, factor 4.

Abraham et al. (1999) studied forward selection and all subsets selection in detail. They showed, by simulating data from several different experiments, that the factors identified as active could change completely if a different fraction was used and that neither of these methods could reliably find three factors which have large effects. However, they concluded that all subsets selection is better than forward selection. Kelly and Voelkel (2000) showed more generally that the probabilities of type-II errors from stepwise regression are high.

The first paper to concentrate on the analysis of data from supersaturated designs was by Westfall et al. (1998). They suggested using forward selection with adjusted $p$-values to control the type-I error rate. They showed how to obtain good approximations to the true $p$-values using resampling methods, but concluded that control of type-I and type-II errors in supersaturated designs is fundamentally a difficult problem. Applying their adjusted $p$-values to the data in Table 6, they found that only the effect of factor 15 is significantly different from zero.

## *3.2 Biased Estimation Methods*

The methods described above all use ordinary least squares to fit several different submodels of the main effects model. Biased estimation methods attempt to fit the full main effects model by using modifications of the least squares method. Lin (1995) suggested using ridge regression, that is, replacing $\boldsymbol{X}'\boldsymbol{X}$ with $\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}$ for some $\lambda$ and then inverting this matrix instead of $\boldsymbol{X}'\boldsymbol{X}$ in the least squares equations. However, he reported that ridge regression seems to perform poorly when the number of factors, $f$, is considerably greater than the number of runs, $n$.

Li and Lin (2002) used a form of penalized least squares with the smoothly clipped absolute deviation penalty proposed by Fan and Li (2001). This method estimates the parameters, $\boldsymbol{\beta}$, by minimizing not the usual residual sum of squares, but

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i'\boldsymbol{\beta})^2 + \sum_{j=1}^{f+1}\phi(\beta_j), \tag{14}$$

where the penalty, $\phi(\beta_j)$, shrinks small estimated effects towards zero. It is defined by its first derivative,

$$\frac{\partial \phi}{\partial \beta} = \lambda \left\{ I(\beta \leq \lambda) + \frac{3.7\lambda - \beta}{2.7\lambda} I(\beta > \lambda) \right\}, \quad \beta \neq 0, \tag{15}$$

and $\phi(0) = 0$, where $I(\cdot)$ is an indicator function and $\lambda$ is a tuning constant chosen from the data by cross-validation. Li and Lin also showed that a good approximation to this method is given by iterated ridge regression. They showed that this method greatly outperformed stepwise variable selection in terms of finding the true model from simulated data. In the data set considered here, it identified 15, 12, 19, and 4 as the active factors.

Comparing the results from different methods of analyzing the data in Table 6, it can be seen that they generally agree on the ordering of effects, namely, 15, 12, 19, 4, and 10, but that they lead to different decisions about which factors should be declared active and which should not.

### *3.3 Bayesian Methods*

A partially Bayesian approach was suggested by Chipman et al. (1997). They used independent prior distributions for each main effect being active. The prior distribution selected for $\beta_j$ was a mixture of normals, namely, $N(0, \tau_j^2)$ with prior probability $1 - \pi_j$ and $N(0, c_j\tau_j^2)$ with prior probability $\pi_j$, where $c_j$ greatly exceeds 1. The prior distribution for $\sigma^2$ was a scaled inverse-$\chi^2$. They then used the Gibbs-sampling-based stochastic search variable selection method of George and McCulloch (1993) to obtain approximate posterior probabilities for $\beta_j$, that is, for each factor they obtained the posterior probability that $\beta_j$ is from $N(0, c_j\tau_j^2)$ rather than from $N(0, \tau_j^2)$. They treated this as a posterior probability that the corresponding factor is active and used these probabilities to evaluate the posterior probability of each model.

Beattie et al. (2002) used the posterior probabilities of factors being active to compare models based on subsets of the candidate factors using the intrinsic Bayes factors of Berger and Pericchi (1996). This approach involves starting with noninformative priors, using a small training sample (part of the data) to obtain posteriors, which are then used as proper priors to obtain Bayes factors from the rest of the data. Because there is no identifiable training set, they used the arithmetic or geometric mean intrinsic Bayes factors over each possible choice of training set. The Bayes factor for comparing two models is interpreted as the ratio of the increase in posterior odds over prior odds for the two models.

## 4 Recommendations and Future Research

Something is wrong! Several methods for analyzing data from supersaturated designs have been proposed, but none of them seem very convincing. Designs are

usually built to optimize the $E(s^2)$ criterion, but this appears to be unrelated to the way in which the data are analyzed. The potential user of supersaturated designs needs to know the answers to three questions.

1. Should supersaturated designs ever be used? If so, in what circumstances? If not, what should be used instead?
2. How should data from supersaturated designs be analyzed and interpreted?
3. How should supersaturated designs be constructed?

In my opinion, we do not know how to answer any of these questions! In this section, each is discussed in turn, although they are all interconnected. The objective is not to answer these questions, but to clarify what research needs to be done to enable them to be answered.

## *4.1 Alternatives to Supersaturated Designs*

If experimenters have a list of $f$ potentially important factors that might affect their process and no prior knowledge about which are more likely to be important, but only sufficient resources to do $n(< f + 1)$ experimental runs, what alternatives to supersaturated designs exist? They might consider any of the following.

1. Abandon the idea of experimentation and seek other scientific or engineering solutions, or abandon or redefine the overall objective of their research program;
2. Experiment with only $n - 1$ factors (a saturated design);
3. Use the group screening methods described in Chapter 9.

All of these are serious competitors to the use of supersaturated designs.

Option 1 should be considered seriously and often this will have been done before any discussion of running an experiment. If the costs of experimentation are greater than the expected gains from improvements discovered by experimentation, then it is obvious that it is better not to run the experiment. This applies to any experiment, not just those using supersaturated designs or those for factor screening. The incomplete knowledge about the benefits of supersaturated designs, however, makes it particularly difficult to relate their expected outcomes to costs.

Selecting only $n - 1$ factors for experimentation is probably what is done most often in practice, the factors being selected by using available knowledge or guessing. Many statisticians' first reaction to such a proposal is very negative, because there is a feeling that evidence from the data must be better than guesswork. One situation in which this is not true is if there is prior knowledge about which factors are more likely to have large effects; see Curnow (1972). If there is real prior knowledge that some factors are more likely to be important, it might be better to use a saturated design in these factors, rather than including many more factors that are not expected to have large effects. Which particular forms of prior knowledge are needed to make a saturated design for $n - 1$ factors better than a supersaturated design for $f$ factors is an interesting area for future research.

Even if there is no prior knowledge about which are likely to be the most important factors, it is not immediately obvious that a supersaturated design is better than a saturated design with a random selection of $n - 1$ factors. Allen and Bernshteyn (2003) give an example for 23 factors in 14 runs where, if the prior probability of each factor being active is 0.4, the supersaturated design has a probability of correctly identifying the active factors which is no greater than randomly declaring a certain number to be active. Therefore the supersaturated design must do worse than randomly picking 13 factors for experimentation. Of course a prior probability of 0.4 of a factor being active contradicts the basic assumption of factor sparsity. Nevertheless, this example gives a warning about the limitations of supersaturated designs. Allen and Bernshteyn showed that with smaller probabilities, the supersaturated designs performed better.

The most obvious competitors to supersaturated designs are group screening designs (Chapter 9), which could be considered as a particular type of supersaturated design with each $s_{ij} = \pm n$ or 0. They are regular Resolution II fractional replicates, whereas supersaturated designs are usually irregular Resolution II fractional replicates. (In this sense Option 2 above is to use a regular Resolution I fractional replicate; see Chapter 1 for a discussion of resolution.) Some research to compare these two types of design is needed. Group screening designs have the advantage that they are very likely to identify any factor with a very large main effect, but the disadvantage that its effect is estimated with correlation 1 with several other factors' effects so that a second stage of experimentation is *always* needed to identify the important individual factors. Supersaturated designs have the advantage that the correlations are always less than 1, but the disadvantage of a more tangled set of correlations which can make it easier to miss an important factor. Interestingly, group screening designs based on orthogonal main effects designs are $E(s^2)$-optimal supersaturated designs.

Personally, I would recommend the use of supersaturated designs, but only in certain circumstances and only with a clear warning about their limitations. Screening is usually assumed to be a way of separating a few factors with nonnegligible effects from many with negligible effects. All the evidence suggests that supersaturated designs are not very good at this. However, what they can do is separate a very small number of factors with very large effects from those with smaller, but perhaps nonnegligible, effects. We might call such factors *dominant* to distinguish them from other factors with nonnegligible effects, which we call *active*.

It might be argued that experimenters will usually have information about dominant factors already and will have set up the process to avoid their having a deleterious effect. However, in this case, we are not in the state of complete ignorance for which supersaturated designs are intended. If we really have a very large number of candidate factors and no reliable knowledge of the likely sizes of their effects, then a first stage in experimentation should concentrate on finding the dominant factors if any exist, so that large improvements in the system under study can be made as quickly as possible, perhaps after further experimentation with these dominant factors. Later stages in experimentation can then be devoted to finding other active factors. The supersaturated design will at least have given

us some idea about the likely sizes of their effects that will be useful for designing further experiments. If no dominant factors are found, experimenters can be reassured that they are not running the process disastrously below optimum conditions and can make a better informed decision about whether further experimentation will be useful to identify the active factors.

Further research is needed to make these recommendations more precise.

## *4.2 Data Analysis*

Most of the methods for analyzing data from supersaturated designs have been adapted from methods tailored for saturated or unsaturated designs. This might be a mistake as supersaturated designs are fundamentally different. Consider from first principles how we should carry out frequentist inference and Bayesian analysis.

If the experimental runs are completely randomized, then randomization theory (see Hinkelmann and Kempthorne, 1994) tells us that least squares gives us unbiased estimators of any pre-chosen set of $n - 1$ linearly independent contrasts among the $n$ combinations of factor levels (treatments). In most factorial experiments the pre-chosen treatment contrasts would be main effects and, perhaps, interactions. However, in supersaturated designs there is no rational basis for choosing a set of $n - 1$ contrasts before the analysis. Any model selection method will lead to selection biases, perhaps large biases, in the estimators of effects. If $\sigma^2$ is assumed known, then we can test the null hypothesis that all $n$ treatment populations have equal means. This would not be of great interest, because even if this null hypothesis were true it would not imply that all main effects are zero, only that a particular set of $n - 1$ linear combinations of treatment means are zero. Of course, in practice, $\sigma^2$ is not known.

So, from a frequentist viewpoint, supersaturated designs do not allow us to carry out any useful estimation or inference. However, estimation and inference are not the objectives of running supersaturated designs. Identifying the dominant factors is the objective. The estimation- and inference-based methods that have been recommended for analysis are used indirectly for this objective, but there is no reason to assume that they should be good for this. It is important to recognize that data analysis from supersaturated designs should be exploratory and not inferential.

What we obtain from a supersaturated design is a set of $n$ observations in which we can try to look for patterns. The patterns of interest are indications that the response is related to the main effects of one or more factors. This does, in fact, make the regression-based methods seem reasonable, though perhaps with some modifications. The idea of fitting all simple linear regressions and picking out the factors with the largest effects does not seem unreasonable. It should identify any dominant factors, although it might also identify several other factors whose effects are highly correlated with a dominant factor's effects. Again, we emphasize that hypothesis testing plays no part in this and the estimates obtained are biased. It should be used only as a method of identifying large effects.

Alternatively, a form of forward selection could be used to try to eliminate effects that look large only because they are highly correlated with a dominant

TABLE 8. Order of including factors and estimates obtained.

| Factor | 15 | 12 | 19 | 4 | 10 | 11 |
|---|---|---|---|---|---|---|
| Effect at inclusion | −53.2 | −22.3 | −24.8 | 22.1 | −9.4 | 8.2 |
| Effect in full model | −70.6 | −25.6 | −29.0 | 21.8 | −10.0 | 8.2 |

factor's effect. However, the standard form of selecting effects for inclusion seems inappropriate. Any effect that is highly correlated with the first effect included will never be included in the model because the correlation makes its standard error so high that the corresponding $t$ or $F$ statistic will always be small. Instead, perhaps we could include in the model the effect with the largest point estimate at each stage. Again, this is not attempting to find the correct model, just to identify the factors that are dominant and give some suggestions about which are most likely to be active. From our data we include factors in the order shown in Table 8. This table also shows the least squares estimates of their effects at inclusion and in the fullest model we have fitted, that with all the six factors listed in Table 8. Similar ideas could be used with stepwise or all subsets selection.

These two methods should allow us to identify any truly dominant factors and give some suggestions about which factors are most likely to be active. More than this cannot really be expected. In the example, we can conclude that factor 15 is the only apparently dominant factor and candidates as active factors are 12, 19, and 4. From the complete set of simple linear regressions, summarized in Table 7, we also note that factor 16 might be active, or even dominant, although probably only if we have incorrectly identified factor 15 as being dominant. These two factors are known from the design to be highly correlated and, although 15 is the stronger candidate to be declared dominant, we would not do so with any great confidence. Further experimentation with as many of these factors as possible might be fruitful. These methods deserve further research.

From a Bayesian perspective, the problem with the standard methods of analysis, Bayesian or not, is that they eliminate all factors whose effects have a moderate posterior probability of being close to zero and select all factors whose effects have a low posterior probability of being close to zero. Instead we should eliminate only those factors whose effects have a high posterior probability of being close to zero and should select those whose effects have a moderate probability of being far from zero, perhaps leaving some in a state between selection and elimination. In other words, current methods eliminate factors with a moderate probability of being inactive, even if they also have a moderate probability of being dominant, whereas they keep factors that are almost certainly active, but not dominant. Instead we should keep all factors with nontrivial probabilities of being dominant, even if we learn little about them from the supersaturated design. Those that seem likely to be active but not dominant might or might not be the subject of further experimentation, depending on costs and resources. In either case, the knowledge gained about them might prove useful for future investigations.

The prior distributions used by Chipman et al. (1997) and Beattie et al. (2002), described in Section 3.3, seem ideally suited for a Bayesian analysis. However,

their use of variable selection and hypothesis testing methods, rather than doing a standard Bayesian analysis, seems unnecessarily complex for the problem of interpreting data from supersaturated designs, for the reasons described above in this section. In particular, after defining the priors for the main effects to be mixtures of normal distributions, the use of the mixing probability parameter as an indicator of the probability of a factor being active seems misleading, because by this definition it is possible for an inactive factor to have a larger effect than an active factor. Allen and Bernshteyn (2003) tried to correct this by using a censored (close to zero) normal distribution for the more widely spread component of the mixture. This is not foolproof and anyway introduces more complexity.

The normal mixture prior does not need to represent a true mixture of two different types of effect, active and inactive. Such a distinction may be convenient for interpretation, but is artificial in the modeling and should be introduced later at the interpretation stage. Instead the use of a normal mixture can be just a convenient way of representing prior beliefs with a heavy-tailed distribution. A scaled $t$, double exponential, or Cauchy prior distribution could be used instead, but the mixture of two normal distributions is more flexible.

After obtaining the posterior distribution, no more formal data analysis is necessary. Rather, at this stage we can work directly with the joint posterior distribution to interpret the results. Plots of the marginal posteriors for each main effect should be useful and we can work out the posterior probability of each factor's main effect being further from zero than some constant, for example, the posterior probability that each factor is dominant and the posterior probability that each factor is active. More research is needed in this area and is under way.

### *4.3 Design*

Given the comments in the two preceding subsections, it is clear that consideration will have to be given to how to design experiments for these forms of analysis. Again, this is an area where more research is needed. However, this does not mean that all previous research has been wasted. All the evidence suggests that designs which are good for several purposes usually fall within the class of $E(s^2)$-optimal designs. However, often some $E(s^2)$-optimal designs are better than others for different purposes. If other criteria are developed for these other forms of analysis, it might be sensible to search for designs that optimize them within the class of $E(s^2)$-optimal designs. This will be useful not only to reduce substantially the amount of computation required for the search, but also to ensure that the designs found are reasonable with respect to other criteria.

## 5 Discussion

Supersaturated designs, and likewise grouping screening designs, provide very little information about the effects of the factors studied, unless they are followed up with further experiments. If it is possible to use a fractional factorial design

of Resolution III or greater, this should be done. However, there are situations in which this is not possible and then using a design that provides very little information is better than obtaining no information. If there is a likelihood that only one experiment will be possible or economic, then a supersaturated design might be better than a group screening design. If we are in this situation, then an $E(s^2)$-optimal supersaturated design will be clearly better than a random balance design or any other known design.

In summary, there are some limited practical situations in which supersaturated designs hold great promise and they are essentially ready for use in practice, although more research is needed to ensure that the best use is being made of resources. Good, although perhaps not optimal, designs are already available for many sizes of experiment. The quick and dirty method of analysis described in Section 4.2 should give enough information to identify dominant factors. For situations where there really is no prior knowledge of the effects of factors, but a strong belief in factor sparsity, and where the aim is to find out if there are any dominant factors and to identify them, experimenters should seriously consider using supersaturated designs.

## *References*

Abraham, B., Chipman, H., and Vijayan, K. (1999). Some risks in the construction and analysis of supersaturated designs. *Technometrics*, **41**, 135–141.

Allen, T. T. and Bernshteyn, M. (2003). Supersaturated designs that maximize the probability of identifying active factors. *Technometrics*, **45**, 90–97.

Beattie, S. D., Fong, D. K. H., and Lin, D. K. J. (2002). A two-stage Bayesian model selection strategy for supersaturated designs. *Technometrics*, **44**, 55–63.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109–122.

Booth, K. H. V. and Cox, D. R. (1962). Some systematic supersaturated designs. *Technometrics*, **4**, 489–495.

Box, G. E. P. (1959). Discussion of Satterthwaite and Budne papers. *Technometrics*, **1**, 174–180.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, New York.

Budne, T. A. (1959). The application of random balance designs. *Technometrics*, **1**, 139–155.

Bulutoglu, D. A. and Cheng, C. S. (2003). Hidden projection properties of some nonregular fractional factorial designs and their applications. *Annals of Statistics*, **31**, 1012–1026.

Bulutoglu, D. A. and Cheng, C. S. (2004). Construction of $E(s^2)$-optimal supersaturated designs. *Annals of Statistics*, **32**, 1162–1678.

Butler, N. A., Mead, R., Eskridge, K. M., and Gilmour, S. G. (2001). A general method of constructing $E(s^2)$-optimal supersaturated designs. *Journal of the Royal Statistical Society B*, **63**, 621–632.

Cela, R., Martínez, E., and Carro, A. M. (2000). Supersaturated experimental designs. New approaches to building and using it: Part I. Building optimal supersaturated designs by means of evolutionary algorithms. *Chemometrics and Intelligent Laboratory Systems*, **52**, 167–182.

Chen, J. and Lin, D. K. J. (1998). On the identifiability of a supersaturated design. *Journal of Statistical Planning and Inference*, **72**, 99–107.

Cheng, C. S. (1997). $E(s^2)$-optimal supersaturated designs. *Statistica Sinica*, **7**, 929–939.

Cheng, C. S., Deng, L. Y., and Tang, B. (2002). Generalized minimum aberration and design efficiency for nonregular fractional factorial designs. *Statistica Sinica*, **12**, 991–1000.

Cheng, C. S. and Tang, B. (2001). Upper bounds on the number of columns in supersaturated designs. *Biometrika*, **88**, 1169–1174.

Chipman, H., Hamada, M., and Wu, C. F. J. (1997). Bayesian variable selection for designed experiments with complex aliasing. *Technometrics*, **39**, 372–381.

Curnow, R. N. (1972). The number of variables when searching for an optimum. *Journal of the Royal Statistical Society B*, **34**, 461–476.

Deng, L. Y., Lin, D. K. J., and Wang, J. N. (1994). Supersaturated design using Hadamard matrix. *IBM Research Report RC19470*, IBM Watson Research Center, New York.

Deng, L.-Y., Lin, D. K. J., and Wang, J. (1996). A measurement of multifactor orthogonality. *Statistics and Probability Letters*, **28**, 203–209.

Deng, L.-Y., Lin, D. K. J., and Wang, J. (1999). A resolution rank criterion for supersaturated designs. *Statistica Sinica*, **9**, 605–610.

Eskridge, K. M., Gilmour, S. G., Mead, R., Butler, N. A., and Travnicek, D. A. (2004). Large supersaturated designs. *Journal of Statistical Computation and Simulation*, **74**, 525–542.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

Hinkelmann, K. and Kempthorne, O. (1994). *Design and Analysis of Experiments*, volume 1. John Wiley and Sons, New York.

Holcomb, D. R. and Carlyle, W. M. (2002). Some notes on the construction and evaluation of supersaturated designs. *Quality and Reliability Engineering International*, **18**, 299–304.

Holcomb, D. R., Montgomery, D. C. and Carlyle, W. M. (2003). Analysis of supersaturated designs. *Journal of Quality Technology*, **35**, 13–27.

Kelly, H. W. and Voelkel, J. O. (2000). Asymptotic-power problems in the analysis of supersaturated designs. *Statistics and Probability Letters*, **47**, 317–324.

Li, R. and Lin, D. K. J. (2002). Data analysis in supersaturated designs. *Statistics and Probability Letters*, **59**, 135–144.

Li, W. W. and Wu, C. F. J. (1997). Columnwise-pairwise algorithms with applications to the construction of supersaturated designs. *Technometrics*, **39**, 171–179.

Lin, D. K. J. (1993). A new class of supersaturated designs. *Technometrics*, **35**, 28–31.

Lin, D. K. J. (1995). Generating systematic supersaturated designs. *Technometrics*, **37**, 213–225.

Liu, M.-Q. and Hickernall, F. J. (2002). $E(s^2)$-optimality and minimum discrepancy in 2-level supersaturated designs. *Statistica Sinica*, **12**, 931–939.

Liu, M. and Zhang, R. (2000a). Computation of the $E(s^2)$ values of some $E(s^2)$ optimal supersaturated designs. *Acta Mathematica Scientia*, **20B**, 558–562.

Liu, M. and Zhang, R. (2000b). Construction of $E(s^2)$ optimal supersaturated designs using cyclic BIBDs. *Journal of Statistical Planning and Inference*, **91**, 139–150.

Liu, Y. and Dean, A. (2004). $k$-circulant supersaturated designs. *Technometrics*, **46**, 32–43.

Lu, X., Hu, W., and Zheng, Y. (2003). A systematical procedure in the construction of multi-level supersaturated design. *Journal of Statistical Planning and Inference*, **115**, 287–310.

Nguyen, N. K. (1996). An algorithmic approach to constructing supersaturated designs. *Technometrics*, **38**, 69–73.

Satterthwaite, F. (1959). Random balance experimentation. *Technometrics*, **1**, 111–137.

Tang, B. and Wu, C. F. J. (1997). A method for constructing supersaturated designs and its $Es^2$ optimality. *Canadian Journal of Statistics*, **25**, 191–201.

Wang, P. C. (1995). Comments on Lin (1993). *Technometrics*, **37**, 358–359.

Westfall, P. H., Young, S. S., and Lin, D. K. J. (1998). Forward selection error control in the analysis of supersaturated designs. *Statistica Sinica*, **8**, 101–117.

Williams, K. R. (1968). Designed experiments. *Rubber Age*, **100**, 65–71.

Wu, C. F. J. (1993). Construction of supersaturated designs through partially aliased interactions. *Biometrika*, **80**, 661–669.

Yamada, S. and Lin, D. K. J. (1997). Supersaturated design including an orthogonal base. *Canadian Journal of Statistics*, **25**, 203–213.

# 9
# An Overview of Group Factor Screening

MAX D. MORRIS

The idea of using a group screening procedure to identify the important or active factors using a small designed experiment was described by Watson (1961) and is now applied in a variety of areas of science and engineering. Watson's work built on the earlier ideas of Dorfman (1943) for screening pooled samples of blood in order to identify diseased individuals using minimal resources. Generalizations and extensions of Watson's technique have been developed by a number of authors who have relaxed some of the stringent assumptions of the original work to make the methods more widely applicable to real problems. An overview of some of the proposed screening strategies is presented, including the use of several stages of experimentation, the reuse of runs from earlier stages, and screening techniques for detecting important main effects and interactions.

## 1 Introduction

In experimental programs involving large numbers of controllable factors, one of the first goals is the identification of the subset of factors that have substantial influence on the responses of interest. There are at least two practical reasons for this. One is the empirical observation that, in many important physical systems, much or most of the variability in response variables can eventually be traced to a relatively small number of factors—the concept of *effect sparsity* (see, for example, Box and Meyer, 1986). When this is true, it is certainly sensible to "trim down" the problem to the factors that are "effective" or "active" before detailed experimentation begins. Even when effect sparsity does not hold, however, it is obvious that careful experimentation involving many factors simply costs more than careful experimentation involving a few. Economic and operational reality may necessitate experimentation in phases, beginning with attempts to characterize the influence of the apparently most important factors while conditioning on reasonable fixed values for other possibly interesting factors. Hence, small *factor screening experiments* are performed not to provide definitive estimates of parameters but to identify the parameters that should be estimated first.

Following Dorfman's (1943) description of an analysis for screening physically pooled samples of blood, the idea of using *group screening* strategies to design factorial experiments was carefully described by Watson (1961) whose analysis is the starting point for much of the subsequent research in this area. As a matter of historical record, Watson credited W. S. Conner as having suggested the idea of group factor screening to him.

This overview begins with a description of Watson's treatment (Section 2) and then briefly discusses a number of modifications and generalizations that have been presented by others (see Kleijnen, 1987, and Du and Hwang, 2000, for additional reviews). Section 3 discusses strategies involving more than two stages and a variety of other issues, including the reuse of runs. Multiple grouping strategies and screening for interactions are discussed in Sections 4 and 5, respectively. The intent of this chapter is not to offer a complete review of all that has been done in the area, but to give the reader a sense of some things that can be done to make group factor screening more applicable in specific situations.

## 2 Basic Group Factor Screening

Watson began his description of the technical problem in the following way.

> Suppose that $f$ factors are to be tested for their effect on the response. Initially we will assume that
>
> (i) all factors have, independently, the same prior probability of being effective, $p\,(q = 1 - p)$,
> (ii) effective factors have the same effect, $\Delta > 0$,
> (iii) there are no interactions present,
> (iv) the required designs exist,
> (v) the directions of possible effects are known,
> (vi) the errors of all observations are independently normal with a constant known variance, $\sigma^2$,
> (vii) $f = gk$ where $g =$ number of groups and $k =$ number of factors per group.
>
> These stringent assumptions are made only to provide a simple initial framework. . . . Actually they are not as limiting as they appear.

The two-level experimental designs that Watson goes on to describe partition the individual factors into the groups referenced in point (vii). In each experimental run, all factors in the same group are either simultaneously at their high values or simultaneously at their low values. In other words, the level of the group factor dictates the level of all individual factors within the group.

Watson's seven points might be restated in the common modern language of linear models (although admittedly with a loss of some elegance and intuitive simplicity) by saying that the collection of $n$ observed responses may be written in matrix notation as

$$\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$ are vectors containing the $n$ responses and the $n$ error variables, respectively, $\mathbf{1}$ is a vector of 1s, $\boldsymbol{\beta}$ is a vector of the $f$ main effect parameters, and $\boldsymbol{X}$ is an $n \times f$ matrix, and where

1. For each element of $\boldsymbol{\beta}$, independently,

$$\begin{aligned} \beta_i &= \Delta/2 \quad \text{with probability } p, \\ &= 0 \qquad \text{with probability } q = 1 - p, \end{aligned}$$

   for some $\Delta > 0$ (Watson's (i) and (ii)),
2. $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ with known $\sigma^2$ (Watson's (vi)),
3. $\boldsymbol{X}$ contains $f$ columns, and has the $(i, j)$th element equal to

   $+1$ if the $j$th individual factor is at its "high" level in the $i$th run
   $-1$ if the $j$th individual factor is at its "low" level in the $i$th run

   (Watson's (iii) and (v)), and
4. $\boldsymbol{X}$ may be written in partitioned form as

$$\boldsymbol{X} = (z_1 \cdot \mathbf{1}' | z_2 \cdot \mathbf{1}' | \ldots | z_g \cdot \mathbf{1}'), \tag{2}$$

   where each $z$ is an $n \times 1$ vector and each $\mathbf{1}'$ is a $1 \times k$ vector, and

$$(\mathbf{1} | \boldsymbol{Z}) = (\mathbf{1} | z_1 | z_2 | \ldots | z_g)$$

   is of full column rank (Watson's (iv) and (vii)).

Section 2.1 contains an example with explicit quantities for several of the expressions listed in these items.

The distinguishing characteristics of basic group factor screening, from the perspective of linear models, are the probabilistic assumption about the value of the parameter vector $\boldsymbol{\beta}$ and the grouped-column restriction on the form of the model matrix $\boldsymbol{X}$. The intent of group screening is to learn important characteristics of the system using fewer (often, far fewer) runs than would be required in a "conventional" experiment. When data are lacking, inference requires that assumptions must be made, and Watson's points (i) and (ii) provide a practical and often reasonable basis for approximate interpretation of the system. In fact, these assumptions are "not as limiting as they appear" (see Watsons's text above); the group screening technique often works quite well in cases where at least some of them are violated.

The generation of the matrix $\boldsymbol{X}$ from $\boldsymbol{Z}$ is a statement of the group design strategy. Physical factors are initially confounded in groups of size $k > 1$ so as to construct artificially a reduced statistical model (which can be estimated based on the desired smaller sample size). By this intentional confounding, the investigator abandons the goal of estimating individual elements of $\boldsymbol{\beta}$, focusing instead on grouped parameters representing estimable functions in the original problem. Practically, the problem can be restated as being

$$\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_g]'$ is a $g \times 1$ vector and $\boldsymbol{Z}$ is an $n \times g$ matrix, and where $\gamma_i$ is the sum of the subset of elements of $\boldsymbol{\beta}$ associated with individual factors in the $i$th group. These grouped parameters can, therefore, be written as

$$\gamma_i = A_i \Delta / 2,$$

where $A_i$ is a binomial random variable with parameters $k$ and $p$, and $A_i$ are independent for $i = 1, 2, 3, \ldots, g$. In the first stage, the focus is on determining which of these grouped parameters is nonzero.

The second stage or "follow-up" experiment is designed to examine only those individual factors included in groups that appear to be effective under the assumptions of the model, those for which $\gamma_i > 0$ based on Watson's assumption (ii). The decisions concerning whether groups are effective could, for example, be based on individual $z$- ($\sigma$ known) or $t$- ($\sigma$ estimated) tests for each grouped parameter. If Watson's assumption (ii) is taken seriously, these would logically be one-sided tests ($H_0 \colon \gamma_i = 0$, $H_A \colon \gamma_i > 0$), but a two-sided test ($H_0 \colon \gamma_i = 0$, $H_A \colon \gamma_i \neq 0$) is more robust against the failure of this assumption and so is often preferred in practice. Factors included in groups that do not appear to be effective are fixed at constant values in the second stage experiment. Watson applied the word "effective" only to individual factors. Here I substitute the more popular term "active", and extend this use to say that a group factor is active if it includes at least one active factor.

Hence, a *two-stage* screening experiment as described by Watson requires a predetermined number $n$ of runs in the first stage and a random number $M$ of runs in the second. The distribution of $M$ and the effectiveness of the screening program (that is, success in labeling individual factors as "active" or "not active") depend on characteristics that can be controlled, at least to some degree:

- value of $k$,
- value of $n$,
- the specific form of the stage 1 decision rule for each group; for example, selection of a significance level $\alpha$ for a $z$- or $t$-test,

and characteristics that cannot:

- value of $p$,
- value of $\Delta$,
- value of $\sigma^2$.

Given values of $p$ and $\Delta/\sigma$, values of $k$, $n$, and $\alpha$ can be selected to produce desirable results, such as small expected sample size or small probability of misclassifying individual factors. As with other statistical design problems, the obvious goals are generally in conflict; smaller $n + E(M)$ generally corresponds to a larger expected number of misclassifications of at least one kind, and so some degree of compromise between expense and performance is required. Watson derived expected values of the number of runs required and the number of factors misclassified, as functions of the parameters. These expressions may be used to evaluate the performance of alternative sampling plans.

## 2.1 Example

To demonstrate the strategy, suppose $f = 50$ experimental factors are to be screened and that a decision is made to do this by forming $g = 7$ groups composed of $k = 7$ factors (as factors 1–7, 8–14, and so on), with the 50th factor added as an "extra" to the 7th group. If $\sigma$ has a known value of 4, say, replicate runs will not be needed and an orthogonal 8-run, 2-level design can be used in stage 1. For example, the 8-run Plackett and Burman (1946) design has design matrix

$$\boldsymbol{Z} = \begin{pmatrix} + & + & + & - & + & - & - \\ - & + & + & + & - & + & - \\ - & - & + & + & + & - & + \\ + & - & - & + & + & + & - \\ - & + & - & - & + & + & + \\ + & - & + & - & - & + & + \\ + & + & - & + & - & - & + \\ - & - & - & - & - & - & - \end{pmatrix}.$$

The expanded $\boldsymbol{X}$ matrix for individual factors would be comprised of the columns of $\boldsymbol{Z}$, each repeated 7 times (or 8 for $\boldsymbol{z}_7$).

Suppose now that the true parameter values are as displayed in Table 1. These values do not exactly correspond to Watson's assumptions; in particular, (ii) is violated. Still, if model (3) is fitted to the data via least squares, estimates of the elements of $\boldsymbol{\gamma}$ will be independent and normally distributed, each with a standard error of $\sqrt{2}$, and with means $E(\hat{\gamma}_1) = 6$, $E(\hat{\gamma}_2) = 0$, $E(\hat{\gamma}_3) = -4$, $E(\hat{\gamma}_4) = 0$, $E(\hat{\gamma}_5) = 0$, $E(\hat{\gamma}_6) = 3$, and $E(\hat{\gamma}_7) = 1$. It is likely that groups 1 and 6 will be declared active, because $z$ statistics based on $\hat{\gamma}_1$ and $\hat{\gamma}_6$ will probably be unusually large. Group 3 would also be detected with high probability, but only if a two-sided test is used; if Watson's working assumption (v) is taken seriously, this would not be the indicated procedure, but the uncertainty associated with many real applications would make two-sided testing an attractive modification. Group 7 might be detected as active, but the probability of this is reduced, for either a one- or two-sided test, by the partial "cancellation" of individual effects of opposite sign

TABLE 1. Individual parameters, grouped as in a first-stage screening experiment, for the example in Section 2.1

| Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 | Group 7 |
|---|---|---|---|---|---|---|
| $\beta_1 = 0$ | $\beta_8 = 0$ | $\beta_{15} = 0$ | $\beta_{22} = 0$ | $\beta_{29} = 0$ | $\beta_{36} = 3$ | $\beta_{43} = 0$ |
| $\beta_2 = 0$ | $\beta_9 = 0$ | $\beta_{16} = -4$ | $\beta_{23} = 0$ | $\beta_{30} = 0$ | $\beta_{37} = 0$ | $\beta_{44} = 0$ |
| $\beta_3 = 2$ | $\beta_{10} = 0$ | $\beta_{17} = 0$ | $\beta_{24} = 0$ | $\beta_{31} = 0$ | $\beta_{38} = 0$ | $\beta_{45} = -2$ |
| $\beta_4 = 0$ | $\beta_{11} = 0$ | $\beta_{18} = 0$ | $\beta_{25} = 0$ | $\beta_{32} = 0$ | $\beta_{39} = 0$ | $\beta_{46} = 0$ |
| $\beta_5 = 0$ | $\beta_{12} = 0$ | $\beta_{19} = 0$ | $\beta_{26} = 0$ | $\beta_{33} = 0$ | $\beta_{40} = 0$ | $\beta_{47} = 0$ |
| $\beta_6 = 0$ | $\beta_{13} = 0$ | $\beta_{20} = 0$ | $\beta_{27} = 0$ | $\beta_{34} = 0$ | $\beta_{41} = 0$ | $\beta_{48} = 3$ |
| $\beta_7 = 4$ | $\beta_{14} = 0$ | $\beta_{21} = 0$ | $\beta_{28} = 0$ | $\beta_{35} = 0$ | $\beta_{42} = 0$ | $\beta_{49} = 0$ |
| | | | | | | $\beta_{50} = 0$ |
| $\gamma_1 = 6$ | $\gamma_2 = 0$ | $\gamma_3 = -4$ | $\gamma_4 = 0$ | $\gamma_5 = 0$ | $\gamma_6 = 3$ | $\gamma_7 = 1$ |

in this group—a violation of Watson's assumption (ii). Mauro and Smith (1982) examined the effect of factor cancellation on screening in the more extreme case where parameters associated with active factors all have the same absolute value, but not the same sign. They presented tables and graphs summarizing a numerical study of performance. Expected degradation in performance is modest when both $p$ and the factor group sizes are small, because the probability is minimal that active factors having main effects of opposite signs mask each other. However, performance degradation becomes a more important issue as either $p$ or the factor group sizes increase. Both the expected number of runs in the second stage of the experiment and the expected number of correctly identified active factors are minimized when, other things being equal, the proportion of active factors with positive main effects equals the proportion with negative main effects. One interesting conclusion from the study is that the factor group size leading to the minimum number of runs is the same whether or not the main effects of the active factors have the same signs.

In the present example, if each of groups 1, 3, 6, and 7 were to be declared active, a follow-up experiment in the associated 29 individual factors might be undertaken to complete the screening process. If an orthogonal design is used, this will require $M = 32$ new runs. This ignores the possibility that some of the original 8 runs might be "reused" which is discussed later. Regardless of the specific orthogonal two-level plan used, the standard error of each parameter estimate will be $\sigma/\sqrt{\text{sample size}} = 4/\sqrt{32}$, so all active factors are likely to be discovered if two-sided testing is used. The expected number of "false positives" will depend on the significance level selected. The total number of runs used here is $n + M = 40$, compared to the 52 that would be required by a minimal orthogonal design for all 50 factors.

## 3 Strategies Involving More Than Two Stages

Most group screening experiments are sequential because the specific form of the second stage design depends upon the analysis of data collected at the first stage. Many other sequential plans are possible. I briefly describe a few strategies that can be viewed as generalizations of (and, in some cases, improvements over) basic group screening.

### *3.1 Multiple Stage Screening and Sequential Bifurcation*

Perhaps the most obvious extension of Watson's basic screening strategy is to grouped factor plans involving more experimental stages. Patel (1962) described multiple stage screening as follows.

- in stage 1, group the $f$ factors into $g_1$ groups of size $k_i = f/g_1$;
- for each group found apparently active at stage 1, the $k_1$ factors are grouped into $g_2$ groups of size $k_2 = f/(g_1 g_2)$;
- ...

- for each group found apparently active at stage $s-1$, the $k_{s-1}$ factors are grouped into $g_s$ groups of size $k_s = f/(g_1 g_2 \cdots g_{s-1} g_s)$;
- for each group found apparently active at stage $s$, all factors are individually examined.

Here, $s$ refers to the number of screening stages not counting the final follow-up in which each of the remaining factors is examined individually. Hence, Watson's description is of a 2-stage procedure characterized by $s = 1$.

Patel offered an analysis based on assumptions that the experimental design used at each stage is of minimal size (and so contains one more run than the number of group factors being investigated) and that $\sigma^2 = 0$. Under these conditions, he showed that values of $g_i$ which minimize the expected number of total runs required are approximately

$$g_1 \approx f p^{s/(s+1)}, \qquad g_i \approx p^{-1/(s+1)}, \quad i = 2, \ldots, s,$$

and that, when these numbers of equal-sized groups are used, the expected number of runs required by this multiple stage procedure is approximately

$$(s+1) f p^{s/(s+1)} + 1.$$

These expressions can, in turn, be used to determine an optimal number of stages for a given value of $p$, at least under the idealized assumptions of the analysis. So, for example, if $p$ were taken to be $6/50$ in the example of Section 2.1, these expressions would yield rounded values of 17 groups and 36 runs, respectively, for the two-stage plans of Watson ($s = 1$). If one additional stage is added ($s = 2$) then the values would be 12 groups for the first stage and 2 groups for the second stage with a total of 37 runs, indicating that the addition of a stage does not improve the expected number of required runs in this case.

Multiple stage screening could also be defined without the requirement of successive splitting of each apparently active group. Instead, the factors included in active groups at one stage could be randomly assigned to smaller groups at the next stage without imposing this constraint.

Bettonvil (1995) discussed a particular form of the multiple group screening idea called *sequential bifurcation*, in which

1. all factors are included in a single group in the first stage, $k_1 = f$, and
2. the factors from an apparently active group at any stage are divided into two subgroups of equal size in the next, $k_{i+1} = k_i/2$.

Chapter 13 reviews sequential bifurcation in more detail.

## 3.2 Orthogonality and Reuse of Runs

The expected number of runs required by a sequential screening plan depends, sometimes heavily, on (1) whether a response value may be used only once in the analysis immediately following the experimental stage in which it is acquired or may be reused in subsequent analyses, and (2) whether the experimental designs

TABLE 2. Example of experimental runs added at each stage and used in each analysis: sequential bifurcation

| | | Factor | | | | | | | | Data used in analysis following stage | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Run | Added at stage | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 1 | 2 | 3 | 4 |
| 1 | 1 | − | − | − | − | − | − | − | − | • | • | • | • |
| 2 | 1 | + | + | + | + | + | + | + | + | • | • | | |
| 3 | 2 | + | + | + | + | − | − | − | − | | • | • | |
| 4 | 3 | + | + | − | − | − | − | − | − | | | • | • |
| 5 | 4 | + | − | − | − | − | − | − | − | | | | • |

used at each stage are required to be orthogonal. Neither Patel (1962) nor Bettonvil (1995) required orthogonality in the designs that they described. This is because each author initially motivated his design for situations in which $\sigma^2 = 0$, and so the usual statistical arguments for precision associated with orthogonality are not relevant. However, Bettonvil allowed the reuse of as many runs as possible from stage to stage, resulting in further reduction in the required number of runs. For example, Table 2 presents a sequence of experimental runs that would be made using sequential bifurcation as described by Bettonvil in an experiment in which there are 8 factors, only the first is active, and no mistakes are made in testing. After stage 1, runs 1 and 2 are used to test all 8 factors as a single group. After stage 2, runs 1, 2, and 3 are used to test group factors (1, 2, 3, 4) and (5, 6, 7, 8); hence runs 1 and 2 are reused. Similarly, following stage 3, runs 1, 3, and 4 are used to test group factors (1, 2) and (3, 4) and, following stage 4, runs 1, 4, and 5 are used to test individual factors 1 and 2. In the analysis following each of stages 2, 3, and 4, the response values of two runs from previous stages are incorporated in the analysis.

In comparison, Table 3 displays a similar description of how Patel's multiple screening would evolve in the same situation. An initial group of all the factors ($g_1 = 1, k_1 = f$) is used, followed in subsequent stages by groups that are half

TABLE 3. Example of experimental runs added at each stage and used in each analysis: multiple stage (Patel •, modified ○)

| | | Factor | | | | | | | | Data used in analysis following stage | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Run | Added at stage | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 1 | 2 | 3 | 4 |
| 1 | 1 | − | − | − | − | − | − | − | − | • | • | • | • |
| 2 | 1 | + | + | + | + | + | + | + | + | • | ○ | | |
| 3 | 2 | + | + | + | + | − | − | − | − | | • | ○ | |
| 4 | 2 | − | − | − | − | + | + | + | + | | • | | |
| 5 | 3 | + | + | − | − | − | − | − | − | | | • | ○ |
| 6 | 3 | − | − | + | + | − | − | − | − | | | • | |
| 7 | 4 | + | − | − | − | − | − | − | − | | | | • |
| 8 | 4 | − | + | − | − | − | − | − | − | | | | • |

TABLE 4. Example of experimental runs added at each stage and used in each analysis: multiple stage with orthogonal plans at each stage and no reuse of runs

| | | Factor | | | | | | | | Data used in analysis following stage | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Run | Added at stage | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 1 | 2 | 3 | 4 |
| 1 | 1 | − | − | − | − | − | − | − | − | • | | | |
| 2 | 1 | + | + | + | + | + | + | + | + | • | | | |
| 3 | 2 | − | − | − | − | − | − | − | − | | • | | |
| 4 | 2 | + | + | + | + | + | + | + | + | | • | | |
| 5 | 2 | + | + | + | + | − | − | − | − | | • | | |
| 6 | 2 | − | − | − | − | + | + | + | + | | • | | |
| 7 | 3 | − | − | − | − | − | − | − | − | | | • | |
| 8 | 3 | + | + | + | + | − | − | − | − | | | • | |
| 9 | 3 | + | + | − | − | − | − | − | − | | | • | |
| 10 | 3 | − | − | + | + | − | − | − | − | | | • | |
| 11 | 4 | − | − | − | − | − | − | − | − | | | | • |
| 12 | 4 | + | + | − | − | − | − | − | − | | | | • |
| 13 | 4 | + | − | − | − | − | − | − | − | | | | • |
| 14 | 4 | − | + | − | − | − | − | − | − | | | | • |

the size of their predecessors. Patel assumed that only the first run (all factors at the low level) is reused at each stage. Hence runs 1, 3, and 4 are used for testing group factors (1, 2, 3, 4) and (5, 6, 7, 8) following stage 2; runs 1, 5, and 6 are used for testing group factors (1, 2) and (3, 4) following stage 3; and runs 1, 7, and 8 are used for testing factors 1 and 2 following stage 4. All eight runs are unique, but all are not strictly necessary for the estimation of the group effects required in the screening strategy. In this example, the analysis of Patel's design can be modified somewhat to allow the reuse of two runs at each stage, as indicated by the open circles in Table 3, but this modification does not change the result of the tests if the assumptions actually hold and there are no random errors in observations.

When $\sigma^2$ is not negligible, the benefits of run reuse and nonorthogonal saturated designs are not so clear-cut. Then, the reuse of runs makes the analysis of performance more complicated because it introduces dependencies between test statistics at each stage. Some duplication of runs (rather than reuse) would allow estimation, or at least a check on the assumed value, of $\sigma^2$. Furthermore, as noted by Watson, when observations include error, many investigators would be more comfortable with the more efficient orthogonal designs. Table 4 shows a sequence of runs for a modified version of multiple stage screening using minimal orthogonal designs at each stage and allowing no reuse of runs from previous stages. This design requires more runs than either Patel's multiple group procedure or Bettonvil's sequential bifurcation, but it provides more precise estimates of group factor parameters in stages 2–4 when there is random error. Furthermore, it allows for the option of blocking to correct for stage effects, or 6 degrees of freedom (after the last stage) for estimating $\sigma^2$ if blocking is not needed. In this example,

the strategy based on orthogonal designs with maximum reuse of runs from stage to stage would be equivalent to the modified multiple stage plan in Table 3.

The decision to use orthogonal plans or to allow nonorthogonal plans, and the decision to allow or disallow the reuse of runs from stage to stage, are related operational issues. They individually and jointly affect the performance of a screening plan and the complexity of calculations required to assess analytically that performance. Depending on characteristics of the application, such as the degree of measurement error and the need to account for block effects in sequential experimentation, either or both may be important considerations.

### *3.3 Stepwise Screening*

Odhiambo and Manene (1987) introduced a stepwise screening plan featuring sequential testing of individual factors after the first (grouped) experiment. After an initial stage as described by Watson, a new experiment is undertaken for each apparently active group in which individual factors are tested one by one until one of them is found to be active. At that point, any remaining factors (not yet individually tested) are tested together as a group, and depending on the result of that test, all are labeled as not active or subjected to further individual examination as following the initial stage. The sequential process of individual tests and group tests, following the discovery of individual active factors, continues until all factors are classified. A schematic of how this might develop in a hypothetical example is given in Figure 1.

This strategy can offer some additional efficiency if some initial groups contain only one active factor, because this factor may be discovered early in follow-up testing and the remaining factors eliminated in one further group test. This occurs in the second and third initial groups in Figure 1. Such efficiency may not necessarily occur, however. For example, identification of the active factors in group 6 (individual factors 26–30) in Figure 1 requires 6 follow-up runs.

Odhiambo and Manene presented a performance analysis of stepwise screening that assumes $\sigma^2 > 0$, where statistical tests are fallible even if all assumptions are correct. They derived expected values of the number of runs required, the number of factors mistakenly classified as active, and the number of factors mistakenly classified as not active, in terms of $p$, $f$, $k$, and the significance level and power of the tests used. These expressions are fairly complicated and are not repeated here, but Odhiambo and Manene also provide simpler approximations that are appropriate for small values of $p$.

## 4 Multiple Grouping Strategies

Sequential group screening methods can lead to substantial test savings; loosely speaking, the more sequential a procedure, in terms of the number of decision points, the greater is the potential for reduction in the expected number of runs required. However, there are settings in which such approaches are operationally impractical, for example, where execution of each run takes substantial time but

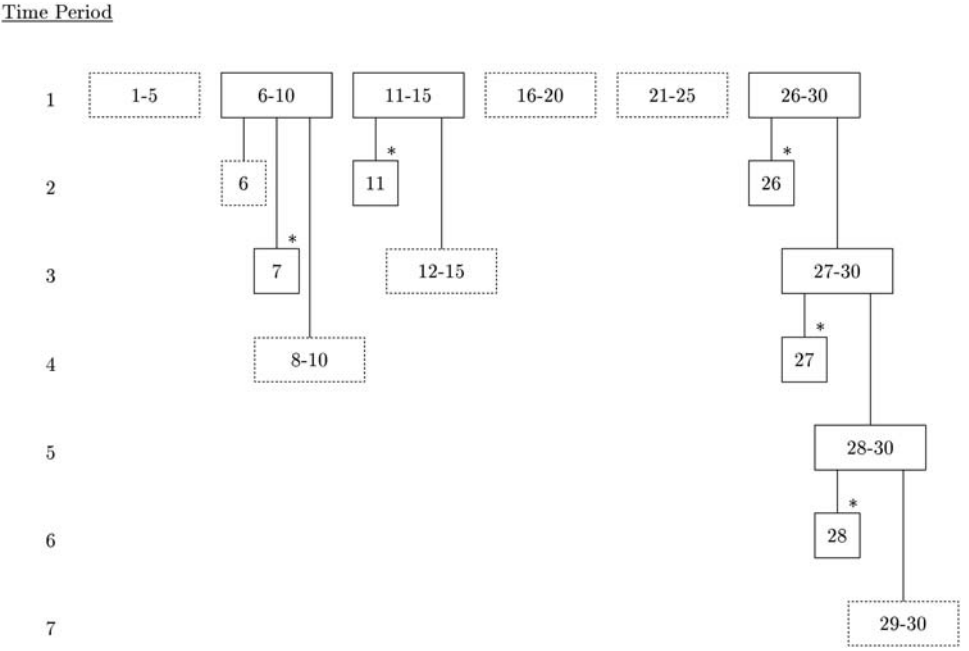



FIGURE 1. Example of analyses and decisions made in a stepwise experiment. The numbers in the boxes refer to factors and each box represents a test of the indicated group or individual factor. Dashed and solid boxes indicate tests in which factors are determined to be not active and active, respectively; asterisks indicate points at which individual active factors are discovered.

many runs can be executed simultaneously as a "batch". Nonsequential procedures based on assigning each object/factor to more than one group were discussed by Federer (1987) for the blood screening problem addressed by Dorfman, and by Morris (1987) for factor screening. These "multiple grouping" methods can in some cases, attain some of the savings of sequential approaches although requiring only one or two temporal sets of tests.

The first, and sometimes only, stage of a multiple grouping screening experiment can be thought of as $r$ simultaneous applications of Watson's original concept, in which the factor groups are defined "orthogonally" in the different applications. Hence $f = 48$ factors might be organized in $g_{(1)} = 3$ type-1 groups of size $k_{(1)} = 16$ factors, and $g_{(2)} = 4$ type-2 groups of size $k_{(2)} = 12$ factors, such that the intersection of any group of type 1 with any group of type 2 contains 4 factors. This arrangement is depicted graphically in Figure 2. The individual factors followed up in the second stage are those for which all types of groups are apparently active. So, for example, if only the first group of type 1 (containing factors 1–16) and the first group of type 2 (containing factors 1–4, 17–20, and 33–36) are declared active, only factors 1, 2, 3, and 4 would be examined in the follow-up experiment. If intersections contain only one factor each, the second stage may be eliminated or used for verification purposes.

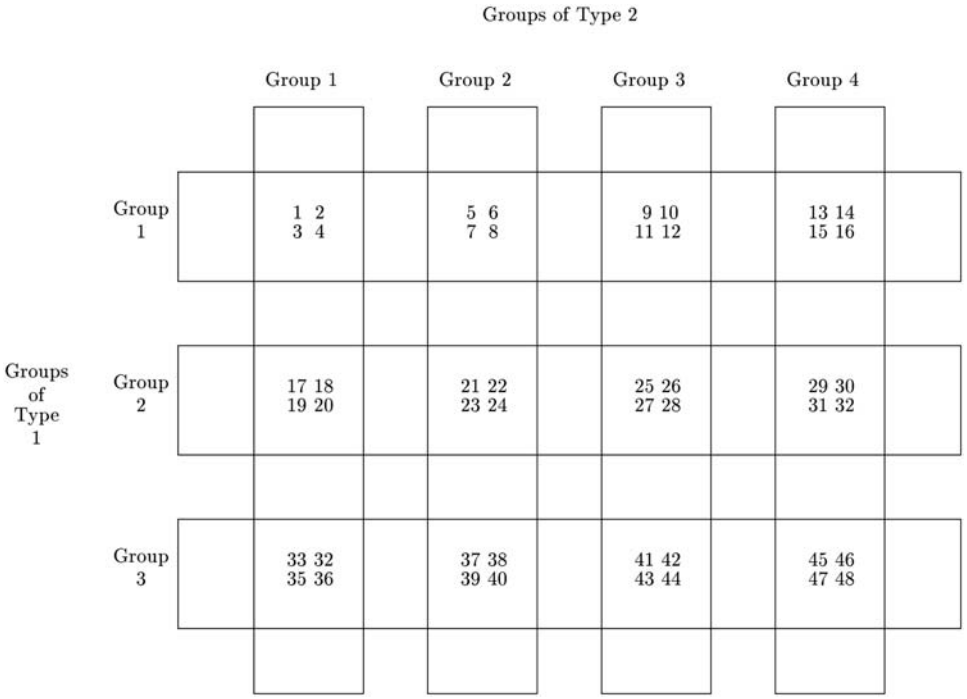



FIGURE 2. Example of individual factor assignments in a multiple grouping screening experiment involving factors labeled $1, 2, \ldots, 48$.

Morris discussed the construction of minimal experimental designs for such procedures, assuming $\sigma^2 = 0$. In most practical applications, performance tradeoffs involving run reuse and orthogonality, as discussed above, would need to be addressed.

## 5 Interactions

All discussion up to this point is predicated fairly seriously on Watson's assumption (iii), that is, the assumption that factors do not interact. However, suppose now that some two-factor interactions do exist so that, as distinct from equation (3), the model is

$$\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \boldsymbol{Z}_2\boldsymbol{\gamma}_2 + \boldsymbol{\epsilon}, \tag{4}$$

where

- $\boldsymbol{Z}_1$ and $\boldsymbol{\gamma}_1$ are as described as $\boldsymbol{Z}$ and $\boldsymbol{\gamma}$ before (representing main effects for factor groups), and
- $\boldsymbol{Z}_2$ is the appropriate model matrix for a set of two-factor interactions, elements of the vector $\boldsymbol{\gamma}_2$.

For experiments in which $\mathbf{Z}_1$ is of full column rank, it is well known that if model (3) is used as the basis for analysis, least squares estimation is biased by the nonzero elements of $\boldsymbol{\gamma}_2$:

$$E(\hat{\boldsymbol{\gamma}}_1) = \boldsymbol{\gamma}_1 + (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Z}_2\boldsymbol{\gamma}_2. \tag{5}$$

This issue is addressed, for instance, in the discussion of model coefficient aliasing in books on response surface analysis such as Myers and Montgomery (2002). It is clear that this aliasing can introduce serious problems into decision processes based upon the realized estimates of model coefficients.

## *5.1 Avoiding Bias Due to Interactions*

In a classic reference, Box and Hunter (1961) noted that "foldover" designs comprised of pairs of runs that are "mirror images" of each other, for example:

$$(+\,+\,-\,-\,-\,+\,-)$$
$$(-\,-\,+\,+\,+\,-\,+),$$

eliminate the aliasing between odd- and even-order effects, and so allow unbiased estimation of main effects even when two-factor interactions exist. Resolution IV main effects plans comprised of foldover run pairs require at least twice as many runs as factors—the operational cost of this benefit. Bettonvil (1993) noted that the sequential bifurcation strategy can be modified to avoid aliasing of main effects with two-factor interactions by adding foldover pairs of runs, rather than individual runs, at each step; similar modifications could certainly be made to the other strategies mentioned here.

## *5.2 Modeling Interactions*

Often interest lies not in simply eliminating the bias from main effect estimates, but also in identifying the interactions that are nonzero. The goal here is screening the effects (main effects and two-factor interactions together) rather than the factors (assuming only main effects are present). Dean and Lewis (2002) and Lewis and Dean (2001) discussed the use of Resolution V designs (which allow estimation of main effects and two-factor interactions) in the group factors in the first stage of a two-stage screening study. These designs use more runs than the Resolution III plans (main effects only) typically used in screening experiments. However, they allow estimation of

- group main effects (the sum of all individual main effects for the group), and
- two-group interactions (the sum of all individual two-factor interactions with one factor in each of two different groups).

Individual two-factor interactions for pairs of factors within the same group are aliased with the intercept and so are not part of any informative estimable combination. In the second stage, the model of interest contains:

- all individual main effects for factors included in an apparently active group main effect,
- all individual two-factor interactions for pairs of factors included in a single group with an apparently active group main effect,
- all individual two-factor interactions for pairs of factors, one each from two groups with an apparently active group interaction, and
- any additional individual main effects required to make the model fully hierarchical.

For example, to screen the main effects and two-factor interactions associated with 20 individual factors, 5 groups of 4 factors each might be formed (say, with factors 1–4 in group 1, and so on). Each $2^{5-1}$ half-replicate associated with the defining relation $I = \pm ABCDE$ is a resolution V design that supports estimation of grouped main effects and two-factor interactions. Suppose that only the main effect associated with group 1 and the two-factor interaction associated with groups 1 and 2 appear to be active in the first stage. Then the individual-factors model used in the second stage would contain:

- an intercept,
- main effects for factors 1–4, because the group 1 main effect is active,
- two-factor interactions for all pairs of factors 1–4, because the group 1 main effect is active,
- two-factor interactions involving one factor from group 1 (1, 2, 3, and 4) and one factor from group 2 (5, 6, 7, and 8), because the interaction for groups 1 and 2 is active, and
- main effects for factors 5–8, so that the model is hierarchical.

The motivating context for this work is robust product design, where each factor is labeled as either a *control* factor or *noise* factor. The distinction between these factors leads to somewhat different effect classification rules and allows the use of group designs of total resolution less than V when some interactions are not of interest. See Lewis and Dean (2001) and Vine et al. (2004) for details, as well as a description of software to evaluate interaction screening designs; the software is available at www.maths.soton.ac.uk/staff/Lewis/screen_assemble/group_screening.html.

## 6 Discussion

The essential characteristic of group screening for factors is the intentional confounding of main effects at various experimental stages, with the aim of reducing the number of runs required to identify those factors that are most important. The number of possible variations on the original theme described in Watson's (1961) paper is nearly limitless. The degree to which runs may be reused, the decision as to whether orthogonal designs should be required at each stage, and modifications to allow consideration of models that include interactions have been briefly considered here.

The formulae for the expected numbers of runs and misclassified factors derived in some of the referenced papers are somewhat complicated, but they are useful in understanding how alternative screening designs and procedures can be expected to perform under simple assumptions. When less stringent assumptions can be made, more elaborate decision rules can be considered. In other circumstances for which classical analysis is difficult, expected performance of competing plans may more easily be evaluated by numerical simulation studies that mimic the screening process. Randomly generated "realities" (such as the number and magnitude of active effects) can be generated, results of each screening strategy/plan applied to the simulated experiment, and those strategies with the best statistical properties (such as smallest expected number of runs or misclassified factors) can be identified. An investigator facing a specific factor screening problem, with specific requirements for replication, blocking, and the possibility that some combination of Watson's working assumptions may be inappropriate, can experiment numerically with the ideas discussed in the literature in order to understand the most relevant performance characteristics of alternative strategies.

## *References*


Bettonvil, B. (1995). Factor screening by sequential bifurcation. *Communications in Statistics—Simulation and Computation*, **24**, 165–185.

Box, G.E.P. and Hunter, J.S. (1961). The $2^{k-p}$ fractional factorial designs, part I. *Technometrics*, **3**, 311–351.

Box, G.E.P. and Meyer, R.D. (1986). An analysis of unreplicated fractional factorials. *Technometrics*, **28**, 11–18.

Dean, A.M. and Lewis, S.M. (2002). Comparison of group screening strategies for factorial experiments. *Computational Statistics and Data Analysis*, **39**, 287–297.

Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics*, **14**, 436–440.

Du, D-Z. and Hwang, F. (2000). *Combinatorial Group Testing and its Applications*, second edition. World Scientific, Singapore.

Federer, W.T. (1987). On screening samples in the laboratory and factors in factorial investigations. *Communications in Statistics—Theory and Methods*, **16**, 3033–3049.

Kleijnen, J.P.C (1987). Review of random and group-screening designs. *Communication in Statistics—Theory and Methods*, **16**, 2885–2900.

Lewis, S.M. and Dean, A.M. (2001). Detection of interactions in experiments on large numbers of factors. *Journal of the Royal Statistical Society B*, **63**, 633–672.

Mauro, C.A. and Smith, D.E. (1982). The performance of two-stage group screening in factor screening experiments. *Technometrics*, **24**, 325–330.

Morris, M.D. (1987). Two-stage factor screening procedures using multiple grouping assignments. *Communications in Statistics—Theory and Methods*, **16**, 3051–3067.

Myers, R. and Montgomery, D. (2002). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, second edition. Wiley, New York.

Odhiambo, J.W. and Manene, M.M. (1987). Step-wise group screening designs with errors in observations. *Communications in Statistics—Theory and Methods*, **16**, 3095–3115.

Patel, M.S. (1962). Group-screening with more than two stages. *Technometrics*, **4**, 209–217.

Plackett, R.L., and Burman, J.P. (1946). The design of optimum multifactorial experiments. *Biometrika*, **33**, 305–325.

Vine, A.E., Lewis, S.M., and Dean, A.M. (2004). Two-stage group screening in the presence of noise factors and unequal probabilities of active effects. *Statistica Sinica*, **15**, 871–888.

Watson, G.S. (1961). A study of the group screening method. *Technometrics*, **3**, 371–388.

# 10
# Screening Designs for Model Selection

William Li

The problem of designing an experiment for selecting a good model from a set of models of interest is discussed in the setting where all factors have two levels. The models considered involve main effects and a few two-factor interactions. Two criteria for the selection of designs for model screening are introduced. One criterion selects designs that allow the maximum number of distinct models to be estimated (estimation capacity). The other maximizes the capability of the design to discriminate among competing models (model discrimination). Two-level orthogonal designs for 12, 16, and 20 runs that are optimal with respect to these criteria are constructed and tabulated for practical use. In addition, several approaches are discussed for the construction of nonorthogonal designs. The chapter includes new results on orthogonal designs that are effective for model discrimination.

## 1 Introduction

An important aspect of designing an experiment is the selection of a design that is efficient for answering the questions of interest to the practitioner. At the initial stage of a project, there is usually a large number of candidate factors and there is a need for screening designs to identify those factors that have an impact on the response. One of the main objectives of a screening design is to build a model that captures the relationship between factors and the response. Because the true model is usually unknown, it is important that the selected design can be used to estimate effectively models within a broad class of possible models.

To accomplish this goal, it is preferable that all models are estimable under the design. The traditional designs, such as minimum aberration fractional factorial designs (see Wu and Hamada, 2000), may not be good choices for this purpose. For example, consider the illustration of Li and Nachtsheim (2000), where the goal of the experiment was to reduce the leakage of a clutch slave cylinder in an automobile. There were four potentially significant factors: body inner diameter, body outer diameter, seal inner diameter, and seal outer diameter. It was believed that the true model should contain all four main effects and was likely to include one or two 2-factor interactions. In this chapter, the main effects of the four factors

and their interactions are represented as 1, 2, 3, 4, 12, 13, 14, 23, 24, 34, rather than by $A$, $B$, and so on, as in Chapter 1.

For this clutch experiment, the experimenters were seeking a design capable of estimating main effects plus any pair of two-factor interactions. Because there are $\binom{4}{2} = 6$ two-factor interactions, the number of models containing main effects and 2 two-factor interactions $\binom{6}{2} = 15$. As pointed out by Li and Nachtsheim (2000), if a resolution IV design with defining relation $I = 1234$ (see Chapter 1) is used, then not all of the 15 possible models can be estimated. For example, the model containing the main effects 1, 2, 3, 4, together with the two-factor interactions 12 and 34, is not estimable because of the aliased pair of interactions $12 = 34$. In fact, only 12 out of the 15 models can be estimated; thus the percentage of all possible models that are estimable for this design is $12/15$, or 80%.

A useful measure of the estimation capability of a design over a class of models is *estimation capacity*, which was introduced by Sun (1993) and then used by Li and Nachtsheim (2000) in a criterion for the construction of model-robust factorial designs. Estimation capacity (EC) is defined as

$$\text{EC} = \frac{\text{number of estimable models}}{\text{total number of possible models}} \tag{1}$$

and is often represented as a percentage. In this chapter, proportions and percentages are used interchangeably.

An ideal screening design would have $EC = 100\%$. In this chapter, such a design is called a *full estimation capacity design*. For the clutch experiment, a full estimation capacity design exists and was constructed by Li and Nachtsheim (2000, Figure 1, page 346). This design is not a minimum aberration design.

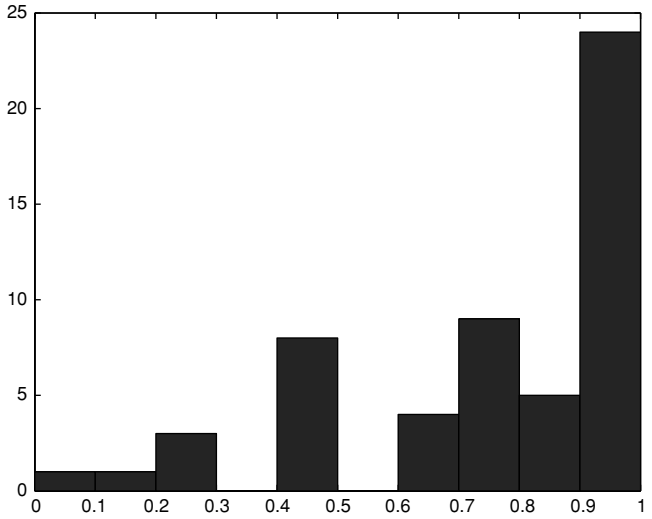



FIGURE 1. Histogram of $EC_2$ values for all 55 designs with 7 factors and 16 runs.

For some numbers of factors and runs, full estimation capacity designs do not exist. The objective is then to find a design that maximizes $EC$—called an *EC-optimal design*. The construction of $EC$-optimal designs is discussed in the first part of this chapter.

When choosing a screening design, the estimability over all possible models is only the first consideration. A further issue needs to be addressed. Even if two models are each estimable, it is not guaranteed that they can be separated in the model analysis procedure. For instance, suppose that, in the clutch experiment, the design with defining relation $I = 1234$ is used, and suppose that the true model contains main effects of factors 1, 2, 3, 4 and interactions 12 and 13. Now consider a competing model that contains main effects and interactions 1, 2, 3, 4, 12, and 24. The two models would produce exactly the same predictions because $13 = 24$. Consequently, a practitioner would not be able to identify the true model and we say that it is *aliased* with another possible model. The *model discrimination capability* of the design is discussed in the second part of the chapter.

The following notation is used in this chapter. Consider a design $d$ with $n$ runs and $f$ factors having two levels each. The design can be represented by an $n \times f$ *design matrix* $\boldsymbol{D} = [x_1, \ldots, x_n]'$, where each row of $\boldsymbol{D}$ has elements $+1$ and $-1$. Each row of $\boldsymbol{D}$ defines the levels at which the factors are observed in the corresponding run and is called a *factor level combination* or *design point*. A linear model for representing the response is

$$y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\beta}$ is a vector of $h$ unknown parameters and the error vector $\boldsymbol{\epsilon}$ has mean 0 and variance $\sigma^2 I_n$. The *model matrix* (or expanded design matrix) is given by

$$\boldsymbol{X} = [\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_n)]', \tag{3}$$

where the functional $\boldsymbol{f}$ indicates which effects are present in the model. For example, if the model consists of all main effects and the intercept term, then, at a given design point, $\boldsymbol{x} = (x_1, \ldots, x_f)$,

$$\boldsymbol{f}'(\boldsymbol{x}) = (1, x_1, \ldots, x_f).$$

As a second example, suppose that the model contains all main effects and all two-factor interactions, then

$$\boldsymbol{f}'(\boldsymbol{x}) = (1x_1, \ldots, x_f, x_1x_2, x_1x_3, \ldots, x_{f-1}x_f).$$

Both $\boldsymbol{D}$ and $\boldsymbol{f}$ are important for the design selection problem. The former denotes the design to be used for the experiment and the latter indicates the underlying model to be fitted. The impact of both $\boldsymbol{D}$ and $\boldsymbol{f}$ is reflected in the model matrix $\boldsymbol{X}$, which plays a key role for comparing designs under many commonly used optimality criteria for design selection. For the linear model (2), the least squares estimate of $\boldsymbol{\beta}$ is given by $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ with corresponding variance–covariance matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$. Many optimality criteria are based on the information matrix $\boldsymbol{X}'\boldsymbol{X}$. For example, a $D$-optimal design maximizes $|\boldsymbol{X}'\boldsymbol{X}|$, and an A-optimal design minimizes *trace* $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

From a practical perspective, the design selection problem amounts to finding the best design matrix $\boldsymbol{D}$ such that the design is optimal with respect do a criterion based on a model $\boldsymbol{f}$. Thus, the design matrix $\boldsymbol{D}$, the model $\boldsymbol{f}$, and the criterion that depends on the model matrix $\boldsymbol{X}$ are three key elements that must be considered. There are several important issues concerning the relationships among these three elements.

Firstly, in the literature, the term *orthogonal design* usually refers to the design matrix, not the model matrix; that is, $\boldsymbol{D}'\boldsymbol{D}$ is $n\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix, but $\boldsymbol{X}'\boldsymbol{X}$ may or may not be an identity matrix. Secondly, the orthogonality of the model matrix depends on the model $\boldsymbol{f}$. If only main effects are present, then $\boldsymbol{X}$ is orthogonal provided that $\boldsymbol{D}$ is orthogonal and that entries in every column of $\boldsymbol{D}$ add to zero. In this case, some optimality criteria that are defined on $\boldsymbol{X}$, such as $D$-optimality, are achieved. From the perspective of optimal designs, orthogonal designs are usually the best choice for main-effect-only models. When the models are more complex, orthogonal designs may or may not be optimal. Thirdly, a design $\boldsymbol{D}$ that is optimal with respect to one model $\boldsymbol{f}_1$ may not be optimal with respect to another model $\boldsymbol{f}_2$.

In this chapter, we discuss the choice of screening designs for model selection via the three elements of the design matrix $\boldsymbol{D}$, the model $\boldsymbol{f}$, and a criterion based on $\boldsymbol{X}$. One important feature about the design screening problem is that the true model is usually unknown. If we denote the set of all possible models that might be fitted by $\mathcal{F} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_u\}$, where $u$ is the number of all possible models, then the optimality criterion for design selection should be based on all possible models, rather than on a specific model in $\mathcal{F}$.

The outline for the remainder of the chapter is as follows. In Section 2, a general framework for constructing efficient screening designs is provided. $EC$-optimal orthogonal designs that are efficient with respect to several model assumptions are investigated in Section 3. The models all include main effects plus a few two-factor interactions, but they vary in which types of two-factor interactions are included. In Section 4, *model-discriminating* orthogonal designs are identified that are constructed to maximize the design's capability of distinguishing between competing models. New results are presented for 12-, 16-, and 20-run designs. Nonorthogonal designs are considered in Section 5 and some concluding remarks are given in Section 6. Throughout the chapter, only two-level designs are considered.

## 2 Selection of Screening Designs

In this section, a general "framework" for selecting and constructing designs is introduced that addresses the model uncertainties at the screening stage of experimentation. The framework is composed of three elements: model, criterion, and candidate designs. To save space, we can denote the list of possibilities for these three elements by

$$(\mathcal{F}, \mathcal{C}, \mathcal{D}), \tag{4}$$

where $\mathcal{F}$ is the set of possible models (also called the *model space*), $C$ consists of one or more criteria for selecting designs, and $\mathcal{D}$ is the set of candidate designs (also called the *design space*).

## 2.1 Choice of Model Space $\mathcal{F}$

In many factorial designs considered in the literature, the objective is to estimate certain factorial effects, assuming that the remaining effects are negligible. However, it is often the case that a small number of effects which, prior to the experiment, were assumed to be negligible, were not actually negligible. In view of this, Srivastava (1975) classified factorial effects into three categories: (i) those effects that are regarded as certain to be negligible, (ii) those effects that are regarded as necessary to be included in the model, and (iii) the remaining effects, most of which are actually negligible, but a few of which may be nonnegligible. Srivastava then proposed a class of designs, called *search designs*, that can estimate all effects of type (ii) and also search for nonnegligible effects in category (iii). The model space for the search design is a rather general one:

$$\mathcal{F} = \{\text{models involving all effects of type (ii)} + \text{up to } q \text{ effects of type (iii)}\}. \tag{5}$$

Sun (1993) and Li and Nachtsheim (2000) considered a more restricted model space, namely,

$$\mathcal{F} = \{\text{models involving all main effects} + \text{up to } q \text{ two-factor interactions}\}. \tag{6}$$

In their work, the main effects are considered to be important and are forced into the model. The designs are then constructed to accommodate as many two-factor interactions as possible in addition to the main effects.

There could also be model uncertainties with respect to main effects. In the context of supersaturated designs (see Chapter 8), Li and Nachtsheim (2001) considered the model space

$$\mathcal{F} = \{\text{any } q \text{ out of } f \text{ main effects}\}. \tag{7}$$

Many other papers on supersaturated designs deal with the model space (7), although implicitly, for example, Wu (1993). Other model spaces considered in the literature include that of Li and Nachtsheim (2001), where the model contains $f_1$ main effects known to be nonnegligible, plus any $q$ out of the remaining $f - f_1$ main effects that may be nonnegligible, and the space of nested linear models of Biswas and Chaudhuri (2002).

## 2.2 Choice of Criteria $\mathcal{C}$

Several criteria for design selection have been proposed in the literature. The choice of criterion depends on the objective of the experiment. Here, criteria for model estimation capacity and model discrimination capability are described.

### 2.2.1 Estimation Capacity

Sun (1993) proposed estimation capacity as a measure of the capability of factorial experimental designs for estimating various models involving main effects and interactions. Motivated by this work, Cheng et al. (1999) considered estimation capacity in the context of minimum aberration designs, and Li and Nachtsheim (2000) proposed a class of designs, called model-robust factorial designs, that maximize the estimation capacity. Some of these designs are discussed in detail in Section 3. Using the notation of Li and Nachtsheim, consider the models that contain all main effects and $q$ two-factor interactions; that is,

$$\mathcal{F}_q = \{\text{all main effects} + \text{exactly } q \text{ two-factor interactions}\}. \tag{8}$$

A design with $f$ factors has $t = \binom{f}{2}$ two-factor interactions, so the number of models in the model space $\mathcal{F}_q$ is $u = \binom{t}{q}$.

Let $e_i(d)$ denote the efficiency of design $d$ when the true model $\boldsymbol{f}_T = \boldsymbol{f}_i$, where $\boldsymbol{f}_i$ is the ith model in some model space. A design $d^*$ is *model robust* for $\mathcal{F}$ if it has maximum average weighted efficiency. We write

$$d^* = \text{argmax} \sum_{i=1}^{u} w_i e_i(d), \tag{9}$$

where $w_i \geq 0$ is the weight assigned to model $\boldsymbol{f}_i$ and $\sum_i w_i = 1$. The weight $w_i$ reflects how likely it is that model $\boldsymbol{f}_i$ is the true model. When there is no prior information on which models are more likely to be the true model, equal weights can be used.

Following the definition of $EC$ in (1), let $EC_q(d)$ denote the fraction of models in $\mathcal{F}_q$ that are estimable using design $d$; that is,

$$EC_q = \frac{\text{number of estimable models in } \mathcal{F}_q}{\text{total number of models in } \mathcal{F}_q}, \tag{10}$$

which is often represented as a percentage. It can be seen that (10) is a special case of (9) when $w_1 = w_2 = \cdots = w_u = 1/u$ and

$$e_i(d) = \begin{cases} 1 & \text{if } \boldsymbol{f}_i \text{ is estimable} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The $EC$ of (1) or, more specifically, the $EC_q$ of (10), provides information on the percentage of models over a model space $\mathcal{F}$ or $\mathcal{F}_q$ that are estimable. However, it provides no information on how efficiently models can be estimated. Using the usual $D$-criterion (see Myers and Montgomery (2002, page 394)) as a measure of the efficiency, Sun (1993) proposed the information capacity ($IC$) criterion, which was later extended by Li and Nachtsheim (2000). Specifically, for model space $\mathcal{F}_q$ in (8), Li and Nachtsheim (2000) defined the information capacity $IC_q$ criterion as maximising

$$IC_q = \frac{1}{u} \sum_{i=1}^{u} n^{-1} |\boldsymbol{X}_i' \boldsymbol{X}_i|^{1/h_i}, \tag{12}$$

where $h_i$ is the number of columns in model matrix $\boldsymbol{X}_i$ for the $i$th model in $\mathcal{F}_q$. The difference between this criterion and the original $IC$ criterion of Sun (1993) is that, in (12), the average is taken over all models in $\mathcal{F}_q$, whereas Sun (1993) took the average over all estimable models in $\mathcal{F}_q$. The $IC_q$ criterion of Li and Nachtsheim (2000) was motivated by the pioneering work of Läuter (1974) and Sun (1993). It is a special case of the general definition of (9) with equal weights and the design efficiency defined as $e_i = n^{-1}|\boldsymbol{X}_i'\boldsymbol{X}_i|^{1/h_i}$. Note that the above discussion on information capacity is also applicable to a more general model space $\mathcal{F}$.

### 2.2.2 Model Discrimination Capability

After establishing that all models in a model space $\mathcal{F}$ are estimable (that is, $EC = 100\%$), a practical issue is to determine which model is the correct model. Suppose that $\boldsymbol{f}_1$ is the true model and that it is estimable under a design $d$. Suppose another model $\boldsymbol{f}_2$ is also estimable under $d$. If the columns of the model matrices of both $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ span the same vector space (that is, the columns of one can be written as linear combinations of the columns of the other), then the experimenter would have no way of determining which model is the correct one. In this case, we say the two models, $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, are *fully aliased* under design $d$.

Srivastava (1975) proposed a measure, called *resolving power*, to evaluate the model discrimination capability of a design. If a design can always identify the true model that contains the type-(ii) effects (as defined in Section 2.1) and up to $q$ type-(iii) effects, then the design is said to have resolving power $q$. Srivastava provided a necessary and sufficient condition for a design to have resolving power $q$, based on the assumption of negligible error variability. The pioneering work of Srivastava (1975) is important in this area and his designs can be useful in practical situations where the error variance is small.

Miller and Sitter (2001) used a different approach based on the probability that the true model can be identified for a given design. Using simulation, their method finds the probability that the true model has the lowest residual sum of squares among candidate models of the same size. Similarly, Allen and Bernshteyn (2003) used simulation to identify supersaturated designs that maximise the probability of identifying active factors.

Because simulation usually takes a large amount of computing time, it may not be practical to use a simulation-based approach to evaluate a large number of designs. A more practical method is to use model discrimination criteria that are based on the model matrix. One such criterion is based on the *subspace angle*, introduced by Jones et al. (2005) In the context of evaluating 18-run orthogonal designs. Consider two models $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ with the corresponding model matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively, and denote the corresponding "hat" matrices by $\boldsymbol{H}_i = \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'$ $(i = 1, 2)$. When two hat matrices are equal to each other for a design $d$, then their predictions $\hat{\boldsymbol{y}}_1 = \boldsymbol{H}_1\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_2 = \boldsymbol{H}_2\boldsymbol{y}$ are the same for all values of the response vector $\boldsymbol{y}$. Therefore, two such models are fully aliased under the design. Equivalently, two models are fully aliased when linear vector spaces $V(\boldsymbol{X}_1)$ and $V(\boldsymbol{X}_2)$, spanned by the columns of the model matrices $\boldsymbol{X}_1$

and $\boldsymbol{X}_2$, are the same. One way of measuring the degree of model aliasing is to evaluate the "closeness" between the two vector spaces $V(\boldsymbol{X}_1)$ and $V(\boldsymbol{X}_2)$ via the subspace angle, which is a generalization of angles between two planes in a three-dimensional Euclidean space. It can be defined as

$$a_{12} = \max_{v_1 \epsilon V(\boldsymbol{X}_1)} \min_{v_2 \epsilon V(\boldsymbol{X}_2)} \arccos(v_1' v_2); \tag{13}$$

that is, for each vector $v_1$ in $V(X_1)$ in turn, we search through $V(X_2)$ to find a vector $v_2$ that gives an angle $a(v_1)$ with the smallest $\cos(v_1' v_2)$. Then $a_{12}$ is the largest of all the angles $a(v_1)$. The criterion of (13) can be easily computed in some software; for example, MATLAB (2004) has a command called SUBSPACE for computing the subspace angle. As noted by Jones et al. (2005), this angle is equivalent to

$$a_{12} = \max_{v_1 \epsilon V(\boldsymbol{X}_1)} \arccos(v_1' \boldsymbol{H}_2 v_1), \tag{14}$$

where $\boldsymbol{H}_2$ is the hat matrix corresponding to $\boldsymbol{X}_2$ and so $\boldsymbol{H}_2 v_1$ represents the projection of $v_1$ onto $\boldsymbol{X}_2$.

It follows from (14) that the subspace angle criterion measures the maximum distance between two fitted values obtained using models $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$. When the subspace angle $a_{12}$ is zero, it follows that two models $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ produce the same fitted values for any design point $x$, and the two models are fully aliased. Another way to explain the subspace angle is that it gives a measure of the amount of new information explained by one model that was not explained by the other model. When the subspace angle is 0, two models explain the same amount of information and are fully aliased. Thus the criterion for design selection is to choose the design that maximises the subspace angle (13) or (14). As an example, consider the design composed of the columns labelled 1–11 of the 12-run Hadamard matrix given in Appendix A. If $\boldsymbol{X}_1$ is the model matrix for the main effects of factors 2 and 3, then $\boldsymbol{X}_1$ contains main effects columns identical to the columns labelled 2 and 3 of the Hadamard matrix. Similarly, if $\boldsymbol{X}_2$ contains the columns labelled 4, 5, and 6, then the subspace angle of the spaces spanned by the columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is $\pi/2 = 1.57$ radians, which is the maximum possible angle.

The model discrimination capability of a design can be measured by the differences between the predictions given by different models. This motivated Jones et al. (2005) to propose two criteria based on the differences of predictions: the expected prediction difference and the maximum prediction difference. The expected prediction difference (EPD) is calculated as follows. Consider two models with model matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and hat matrices $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$. Then, for any response $\boldsymbol{y}$, the difference between two predictions is given by $\hat{\boldsymbol{y}}_1 - \hat{\boldsymbol{y}}_2 = (H_1 - H_2)\boldsymbol{y}$. The expected prediction difference measures the average distance between two fitted values over all possible normalised responses, where "normalised" means that the response vector $\boldsymbol{y}$ is scaled so that $\boldsymbol{y}'\boldsymbol{y} = 1.0$. For any two model matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$,

$$\text{EPD}(\boldsymbol{X}_1, \boldsymbol{X}_2) = E((\hat{\boldsymbol{y}}_1 - \hat{\boldsymbol{y}}_2)'(\hat{\boldsymbol{y}}_1 - \hat{\boldsymbol{y}}_2) | \boldsymbol{y}'\boldsymbol{y} = 1) = E(\boldsymbol{y}'\boldsymbol{D}y | \boldsymbol{y}'\boldsymbol{y} = 1), \tag{15}$$

where $\boldsymbol{D} = (\boldsymbol{H}_1 - \boldsymbol{H}_2)^2$. From the definition of (15), EPD = 0 if and only if two

models are fully aliased. A design with good model discrimination capability should maximise (15). The expected prediction difference can be computed easily because it can be shown that EPD $= 1/n\ trace(\boldsymbol{D})$. (For more details, see Jones et al., 2005.)

Both the subspace angle and the expected prediction difference can be used to evaluate the model discrimination capability of a design $d$ over a model space $\mathcal{F}$. In Section 4, the selection of orthogonal designs using criteria based on these measures is discussed.

## 2.3 Choice of Candidate Designs $\mathcal{D}$

Both orthogonal designs and nonorthogonal designs (see Chapter 7) can be used as screening designs for model selection. However, different approaches are taken for the construction of efficient screening designs within these two classes of candidate designs. To find optimal orthogonal designs, we can take advantage of the existing orthogonal designs that are available in the literature. The procedure would simply involve evaluating all existing designs and then selecting the best design(s) with respect to a pre-specified criterion. This approach works particularly well if, for given number of runs $n$ and number of factors $f$, all orthogonal designs are available. Then the global optimal design can be found by exhaustive evaluation for any given criterion.

In some situations, a small sacrifice in orthogonality may result in gains for another criterion that is considered to be more important for screening purposes, in which case nonorthogonal designs may be more desirable. Almost always, it is impractical to evaluate all nonorthogonal designs of a given size. Thus, a different approach is needed to find efficient nonorthogonal designs.

### 2.3.1 Orthogonal Designs

Orthogonal designs have been among the most commonly used designs for several decades. In the literature, the term "orthogonal design" usually refers to a design for which the design matrix (not the model matrix) has orthogonal columns. Such designs are efficient for main-effect-only models. Most of the orthogonal designs discussed in the literature are projections of the Hadamard matrix (see Chapter 7). A Hadamard matrix $\boldsymbol{H}$ of order $n$ is an $n \times n$ matrix of $+1$s and $-1$s such that $\boldsymbol{H}'\boldsymbol{H} = n\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. Construction methods of Hadamard matrices have been discussed by several authors in the literature (see, for example, Hall, 1967).

Not all orthogonal designs are projections of Hadamard matrices. Sun et al. (2002) used a sequential algorithm to obtain the complete catalogue of orthogonal designs for $n = 12$, 16, and 20. They found that all 12-run and 16-run designs are indeed projections of Hadamard matrices, but some of the 20-run orthogonal designs are not projections of the 20-run Hadamard matrices.

Table 1 lists the numbers of nonisomorphic 12-, 16-, and 20-run orthogonal designs. (Two designs are called *isomorphic* if one design can be obtained from another by relabelling runs, relabelling factors, and exchanging factor levels.) For

TABLE 1. Numbers of nonisomorphic 12-, 16-, and 20-run orthogonal designs for $2 \leq f \leq 19$ factors

| $f$ | 12-Run | 16-Run | 20-Run |
|---|---|---|---|
| 2 | 1 | 1 | 1 |
| 3 | 1 | 3 | 3 |
| 4 | 1 | 5 | 3 |
| 5 | 2 | 11 | 11 |
| 6 | 2 | 27 | 75 |
| 7 | 1 | 55 | 474 |
| 8 | 1 | 80 | 1603 |
| 9 | 1 | 87 | 2477 |
| 10 | 1 | 78 | 2389 |
| 11 | 1 | 58 | 1914 |
| 12 | | 36 | 1300 |
| 13 | | 18 | 730 |
| 14 | | 10 | 328 |
| 15 | | 5 | 124 |
| 16 | | | 40 |
| 17 | | | 11 |
| 18 | | | 6 |
| 19 | | | 3 |

instance, it shows that for an experiment with 6 factors, there are two 12-run, twenty-seven 16-run, and seventy-five 20-run orthogonal designs. Table 1 also shows there is one 12-run designs with 11 factors, five 16-run designs with 15 factors, and three 20-run designs with 19 factors. These (saturated) orthogonal designs can be obtained by deleting the column of all 1s from a Hadamard matrix. Thus, the results shown in Table 1 confirm that there is a unique 12-run Hadamard matrix, five 16-run Hadamard matrices, and three 20-run Hadamard matrices. For the convenience of readers, the 12-run and 16-run Hadamard matrices are provided in Appendices A and B. In Appendix C, some 12- and 16-run designs that are efficient for model screening purposes are listed. Complete catalogues of designs are available on the author's Web site: http://www.csom.umn.edu/∼wli. In Sections 3 and 4, these designs are evaluated in terms of maximum estimation capacity and maximum model discrimination criteria.

### 2.3.2 Nonorthogonal Designs

For main-effect-only models, orthogonal designs are usually preferred because main effects can be estimated efficiently and interpreted easily due to the orthogonality between columns in the design matrix. When the model contains interactions as well as main effects, however, a greater concern may be the orthogonality between columns in the model matrix rather than just the orthogonality of columns in the design matrix.

Several authors have recommended the use of nonorthogonal designs for model screening. Li and Nachtsheim (2000) proposed model-robust factorial designs to estimate a class of models containing main effects plus several two-factor interactions. In their work, the model space is the one defined in (6) and the criteria are to maximise $EC_q$ and $IC_q$ defined in (10) and (12). The candidate design space $D$ consists of all balanced designs in which each column has an equal number of $+1$s and $-1$s. This property has recently been called mean orthogonal by Liu and Dean (2004). Balanced designs are often preferred for two reasons. First, an imbalance of $+1$s and $-1$s within a column implies that the estimated factorial effect is correlated with the estimated grand mean, resulting in a lower efficiency of estimation. Secondly, if the level of a factor that produces the better response value happens to be assigned in the design so that it occurs in only a small number of runs, then the experiment may not provide adequate information on the factor main effect. Among the papers that recommended use of nonorthogonal designs both balanced and unbalanced designs have been considered. For example, Miller and Sitter (2005) proposed a class of balanced designs with the aim of maximising the probability that the true model can be identified. Allen and Bernshteyn (2003) considered supersaturated designs of both types that maximise the probability of identifying the important factors.

## 3 Orthogonal Designs Optimal Under Estimation Capacity

In this and the next sections, it is shown that only a few of the orthogonal designs for $n = 12$, 16, and 20 runs, discussed in Section 2.3, are appropriate for the screening purpose of model selection. In this section, the maximum-$EC_q$ criterion, defined through (10), is used to evaluate these designs. Specifically, designs are ranked by the sequential maximization of $(EC_1, EC_2, \ldots)$. The "best" designs are called $EC$-optimal designs. Section 3.1 focuses on the model space defined in (8). Then, in Section 3.2, $EC$-optimal designs for several alternative model spaces are introduced.

### *3.1 EC-Optimal Orthogonal Designs*

In this section, results on optimal orthogonal designs for the model space $F_q$, defined in (8), are introduced. The framework (see Section 2) is given by:

$$(\mathcal{F}_q \text{ as in (8)}, (\max EC_1, \max EC_2, \ldots) \text{ criterion, orthogonal designs}). \quad (16)$$

Optimal designs in the framework (16) are called *model-robust orthogonal designs* with $n$ runs and $f$ factors. For instance, the 16-run six-factor model-robust orthogonal design refers to the optimal orthogonal design, selected from all 27 nonisomorphic 16-run designs with six factors, under the $(\max EC_1, \max EC_2, \ldots)$ criterion. In the framework of (16), the model-robust orthogonal design is obtained for each (n,f). Alternatively, given $n$ and $f$, we can obtain a series of optimal

designs for $q = 1, 2, \ldots$ based on $(EC_q, IC_q)$ of (10) and (12). This approach was taken by Li and Nachtsheim (2000), whose method and results are also mentioned in this chapter.

### 3.1.1 12-Run Model-Robust Orthogonal Designs

All 12-run orthogonal designs are projections of the unique 12-run Hadamard matrix (see Appendix A), which is also known as the 12-run Plackett–Burman design (see, also, Chapter 7). For each of $f = 5$ and 6 factors, there are two nonisomorphic designs. For other numbers of factors, $f$, all selections of $f$ columns result in a set of designs that are isomorphic under row, column, and level permutations; that is, there a unique $f$-factor nonisomorphic design. Lin and Draper (1992) first considered the projection properties of 12-run Plackett–Burman designs. They found that for $f = 5$, the two nonisomorphic projections can be obtained from the columns (1, 2, 3, 4, 5) and (1, 2, 3, 4, 10) of the Plackett–Burman design in Appendix A and, for $f = 6$, the two nonisomorphic projections can be obtained from the columns (1, 2, 3, 4, 5, 6) and (1, 2, 3, 4, 5, 7) of the Plackett–Burman design.

The 12-run designs generally have high $EC$ values. Li and Nachtsheim (2000, Table 3) reported the $EC$ values for designs with $5 \leq f \leq 9$ factors and $q = 1, 2, \ldots, 5$ two-factor interactions in the model. When $q = 1$ (only one two-factor interaction in the model), all designs have full estimation capacity; that is, all models containing main effects plus one two-factor interaction are estimable. When $q = 2$, all the designs given by Li and Nachtsheim for $f \leq 6$ are full estimation capacity designs, and the designs with $f = 7$, 8, and 9 factors have, respectively, $EC = 0.914$, 0.778, and 0.514. For $q \geq 3$, the $EC$ values are between 0.409 and 1.0. The $IC$ values are also reported in Li and Nachtsheim (2000, Table 3). They range from 0.298 (for $f = 8$ and $q = 3$) to 0.944 (for $f = 5$ and $q = 1$).

For $f = 5$ and 6 factors, two nonisomorphic designs exist. The two designs for $f = 5$ have similar $EC$ and $IC$ values. However, the two designs for $f = 6$ are quite different in terms of the $(EC, IC)$ values. The first design, which was reported by Li and Nachtsheim and constructed from columns (1, 2, 3, 4, 5, 6) of the Hadamard matrix, is much better than the second design. The $EC_q (q = 1, 2, 3, 4)$ values for the first design are (1.00, 1.00, 1.00, 0.98), compared with (1.00, 0.86, 0.60, 0.30) of the second design. The $IC_q (q = 1, 2, 3, 4)$ values are (0.93, 0.86, 0.79, 0.68) and (0.93, 0.76, 0.54, 0.32), respectively.

### 3.1.2 16-Run Model-Robust Orthogonal Designs

All 16-run orthogonal designs are projections of the five Hadamard matrices given in Appendix B. As shown in Table 1, the numbers of $f$-factor designs with 16 runs can be reasonably large. These designs may perform very differently in terms of estimation capacity. Figure 1 displays a histogram for the $EC_2$ values for the 55 designs for $f = 7$ factors. The variability in the $EC_2$ values is quite considerable, from a minimum of 0.0 to a maximum of 1.0. Similar patterns are found for other $f$ and $q$ values. Thus, not all orthogonal designs are suitable for model screening purposes.

TABLE 2. Design indices for 16-run $EC$-optimal orthogonal designs over four model spaces of Li and Nachtsheim (2000) (for model-robust orthogonal designs), Bingham and Li (2002) (for model-robust parameter designs), Addelman (1962), and Sun (1993)

| $f$ | Model-robust orthogonal designs | Model-robust parameter designs | | | Addelman $f_1$ | | | | Sun $f_2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $q=1$ | $q=2$ | $q=3$ | 4 | 5 | 6 | 7 | 2 | 3 | 4 |
| 6 | 26 | 5 | 13,19 | 26 | 5 | 19 | — | — | 19 | 3 | — |
| 7 | 52 | 6 | 49,32 | 52 | 6 | 50 | 32 | — | 50 | 50 | — |
| 8 | 68 | 6 | 68 | 68 | 6 | 6 | 68 | 68 | — | 68 | 68 |
| 9 | 85 | 73,85 | 76 | 76 | — | 4 | 71 | 71 | — | — | — |
| 10 | 66 | 76 | 66 | 66 | — | — | — | — | — | — | — |

Some of the 16-run designs summarised in Table 1 are *regular* designs (having defining relations, see Chapter 1). For a regular design, a factorial effect is either independent or fully aliased with each other factorial effect. In contrast, a *nonregular* design (having no defining relation) allows partial aliasing between effects. In the past, regular designs have been the most often used but, in recent years, there has been a surge of interest in nonregular designs. Wu and Hamada (2000) gave a comprehensive review of nonregular designs. In the context of model screening, nonregular designs usually perform much better than regular designs. For example, consider the case of $f = 7$ factors. There are 55 nonisomorphic designs, six of which are regular designs. (The numbers of regular 16-run designs with $f$ factors can be found in, for example, Li et al. 2003, Table 2). For $q = 2$, the estimation capacities for these six regular designs have range (0.00, 0.90) with an average estimation capacity of 0.35. In comparison, the $EC$ range for the remaining 49 nonregular designs is (0.23, 1.00) with average $EC$ of 0.81. Similar results are also observed for other $f$ and $q$ values.

In Table 2, efficient 16-run orthogonal designs for various criteria and model spaces are summarized. In the second column, design indices for the 16-run model-robust orthogonal designs based on the framework of (16) are listed. The designs are obtained by the sequential maximization of $(EC_1, EC_2, \ldots)$. For the reader's convenience, selected 16-run designs, which are efficient for model estimation or model discrimination for $f = 6, \ldots, 10$ factors, are given in Appendix C where the design index number corresponds to that of Sun et al. (2002); the designs are also available on the author's Web site (http://www.csom.umn.edu/∼wli). For example, column 2 of Table 2 shows that the model-robust orthogonal design for $f = 6$ factors is Design 26. This design, denoted in Appendix C by 26 = III(2,4,8,10,12,15), is obtained by using columns 2, 4, 8, 10, 12, and 15 of Hall's design H16.III in Appendix B. The results given in the remaining columns of Table 2 are discussed in Section 3.2.

### 3.1.3 20-Run Model-Robust Orthogonal Designs

The complete catalogue for 20-run orthogonal designs was not available in the literature until recently. Sun et al. (2002) constructed all nonisomorphic 20-run

TABLE 3. Summary of properties of 20-run model-robust orthogonal designs; $p_q$ denotes the percentage of designs for which $EC_q = 100\%$; ave($EC_q$) is the average $EC_q$ over all designs with $f$ factors

| $f$ | $p_1(\%)$ | ave($EC_1$) | $p_2(\%)$ | ave($EC_2$) | $p_3(\%)$ | ave($EC_3$) | $p_4(\%)$ | ave($EC_4$) |
|---|---|---|---|---|---|---|---|---|
| 3 | 66.7 | .677 | 66.7 | .677 | 66.7 | .677 | — | — |
| 4 | 100 | 1.000 | 100 | 1.000 | 100 | 1.000 | 100 | 1.000 |
| 5 | 100 | 1.000 | 100 | 1.000 | 100 | 1.000 | 63.6 | .994 |
| 6 | 100 | 1.000 | 100 | 1.000 | 100 | 1.000 | 40.0 | .997 |
| 7 | 100 | 1.000 | 100 | 1.000 | 100 | 1.000 | 4.2 | .998 |
| 8 | 100 | 1.000 | 99.7 | 1.000 | 99.6 | 1.000 | 65.6 | .999 |
| 9 | 100 | 1.000 | 97.1 | 1.000 | 96.7 | 1.000 | 77.3 | .999 |
| 10 | 100 | 1.000 | 77.6 | .999 | 77.5 | .998 | 56.0 | .995 |
| 11 | 100 | 1.000 | 8.3 | .998 | 7.6 | .993 | 0 | .985 |
| 12 | 100 | 1.000 | 0 | .994 | 0 | .982 | 0 | .961 |
| 13 | 100 | 1.000 | 0 | .958 | 0 | .957 | 0 | .906 |
| 14 | 100 | 1.000 | 0 | .967 | 0 | .900 | 0 | .775 |
| 15 | 100 | 1.000 | 0 | .922 | 0 | .771 | 0 | .489 |
| 16 | 100 | 1.000 | 0 | .828 | 0 | .501 | — | — |
| 17 | 100 | 1.000 | 0 | .605 | — | — | — | — |
| 18 | 100 | 1.000 | 0 | — | — | — | — | — |

designs using an algorithmic approach and obtained $EC$-optimal designs. The numbers of nonisomorphic 20-run designs with $f$ factors ($f = 2, \ldots, 19$) are given in Table 1.

We use the model space in (8) for $q = 1, 2, 3, 4$. In Table 3, two measures are recorded for each value of the number, $f$, of factors and the number, $q$, of two-factor interactions in the model. These measures are the percentage of 20-run designs for which $EC_q = 1.00$, denoted by $p_q$, and the average $EC_q$ over all $f$-factor designs. (In this section, we use a proportion for $EC$ to avoid possible confusion with the percentage $P_q$.) For models with $q = 1$ two-factor interaction, almost all designs have estimation capacity $EC = 1.00$. The only exception is a design with $f = 3$ factors, which consists of five replicates of a regular four-run resolution III design. The defining relation of this design is $I = 123$, and thus $EC_q = 0$ for $q = 1$, 2, and 3 two-factor interactions. When $q = 2$ and 3, $p_q = 100\%$ for $4 \le f \le 7$, indicating that all f-factor designs with 20 runs for $f = 4, \ldots, 7$ have $EC_q = 1.00$; when $q = 4$, $p_4 = 100\%$ only for four-factor designs. In many cases where $p_q < 100\%$, the average $EC$ value is very high. For instance, for $f = 7$ and $q = 4$, only 4.2% of the seven-factor designs have $EC = 1.00$. But the average $EC$ of all designs is 0.998, which indicates that the $EC$ values of the remaining 95.8% of the designs are actually close to 1.00.

In summary, Table 3 shows that most 20-run designs have good $EC$ values for $q = 1$, 2, 3, and 4. However, the estimation capacity should not be used as the only criterion for selecting screening designs. Another useful criterion is the measure of model discrimination capabilities. It is shown in Section 4 that only a small fraction of 20-run orthogonal designs have good model discrimination properties.

## 3.2 EC-Optimal Designs for Alternative Model Spaces

The $EC$-optimal designs discussed in the previous section are based on the model space defined in (8), in which models contain main effects plus $q$ two-factor interactions. Other model spaces have also been considered in the literature. In this section, we introduce the $EC$-optimal designs under several alternative assumptions on the model space.

### 3.2.1 Robust Parameter Designs

Robust parameter designs are used to identify the factor levels that reduce the variability of a process or product (Taguchi, 1987). In such experiments, the dispersion effects, which can be identified by examination of control-by-noise interactions (see Chapter 2), are particularly important and hence the models of primary interest are those that contain at least one control-by-noise interaction. This motivated Bingham and Li (2002) to introduce a model ordering in which models are ranked by their order of importance as follows.

$$\mathcal{F}_1 = \{\text{main effects} + q \text{ two-factor interactions (among which at least one is a control-by-noise interaction)}\}, \quad (17)$$

and

$$\mathcal{F}_2 = \{\text{main effects} + q \text{ two-factor interactions (among which none is a control-by-noise interaction)}\}. \quad (18)$$

Let $EC_1$ and $EC_2$ denote the estimation capacities for the models in $F_1$ and $F_2$, respectively. Denote the information capacity for models in $F_1$ by $IC_1$. Then the model-robust parameter designs proposed by Bingham and Li (2002) sequentially maximise $(EC_1, EC_2, IC_1)$. The design indices for model-robust parameter designs are provided in column 3 of Table 2. For example, the table shows that the model-robust parameter design for $f = 7$ and $q = 1$ is Design 6. According to Appendix C, this design is obtained by using columns 1, 2, 4, 7, 8, 11, and 13 of Hadamard matrix H16.1 in Appendix B. Table 2 shows that, for the same $f$, the model-robust parameter designs may be different for different numbers $q$ of interactions to be estimated. Consider, for example, the case in which $f = 6$. When $q = 1$, the model-robust parameter design is Design 5; but for $q = 2$, the model-robust parameter designs are Designs 13 and 19.

### 3.2.2 Model Spaces Considered by Addelman and Sun

An alternative model space considered by Addelman (1962) is

$$\mathcal{F} = \{\text{main effects} + q \text{ two-factor interactions among } f_1 \text{ specific factors}\}. \quad (19)$$

This model space is appropriate if prior experience indicates that two-factor interactions are likely to be present among only $f_1$ out of $f$ factors. In addition to (19), Addelman (1962) discussed two other model spaces, but the details are omitted

here. Sun (1993) added a new model space to Addelman's work:

$$\mathcal{F} = \{\text{main effects} + q \text{ two-factor interactions between } f_1 \text{ specific factors and the remaining } f - f_1 \text{ factors}\}. \tag{20}$$

This model space can be particularly useful for robust parameter designs, assuming that the nonspecified interactions are negligible. If $f_1$ and $f_2$ denote the numbers of control factors and noise factors, respectively, then the interactions between $f_1$ control factors and $f_2$ noise factors are usually considered to be more important than other two-factor interactions. The $EC$-optimal designs among 16-run orthogonal designs for the models spaces in (19) and (20) were found by Sun (1993). We list the corresponding optimal design indices in columns 4 and 5 of Table 2. With the design index number given in Table 2, the exact design can be found in Appendix C.

## 4 Model-Discriminating Designs

The $EC$-optimal designs discussed in Section 3 maximise the $EC$ criterion. If they are full estimation capacity designs, then all models are estimable. However, this does not imply that such a design will allow the models to be distinguished or separated from each other. In this section, orthogonal designs are further characterised according to their model discrimination capabilities, using the subspace angle (SA) of (13) and expected prediction difference (EPD) of (15). New results are given concerning the discriminating capabilities of 12-, 16-, and 20-run orthogonal designs and recommended designs are tabulated.

Suppose the model space $\mathcal{F}$ contains $u$ models. For each pair of models, the SA and EPD values can be computed. Then the model discrimination capability of a design can be measured by considering the SA and EPD values over all pairs of models, resulting in four measures (Jones et al., 2005):

$$\min \text{SA} = \min_{\boldsymbol{f}_i, \boldsymbol{f}_j \in \mathcal{F}} a_{ij}, \tag{21}$$

$$\text{ave SA} = \frac{1}{u} \sum_{i=1}^{u} a_{ij}, \tag{22}$$

$$\min \text{EPD} = \min_{\boldsymbol{f}_i \boldsymbol{f}_j \in \mathcal{F}} \text{EPD}, \tag{23}$$

$$\text{ave EPD} = \frac{1}{u} \sum_{i=1}^{u} \text{EPD}, \tag{24}$$

where $a_{ij}$ is defined in (13) and (14), and where EPD is defined in (15). In all four measures, larger values indicate better model discrimination properties. Thus, the measures should all be maximised.

In a design for which the minimum subspace angle is zero, there exists at least one pair of models, say $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$, that are fully aliased. Then at any design point $\boldsymbol{x}$, two fitted values $\boldsymbol{f}_i(\boldsymbol{x})$ and $\boldsymbol{f}_j(\boldsymbol{x})$ are identical. Thus, it is important that the min SA of (21) be greater than zero. A design satisfying this condition is called a *model-discriminating design*. It can easily be proved that min SA is zero if and only if min EPD of (23) is zero. Thus, a modeldiscriminating design is a design that satisfies

$$\text{min SA} > 0, \tag{25}$$

or, equivalently,

$$\text{min EPD} > 0. \tag{26}$$

A necessary condition for a design to be model-discriminating is that all models in $\mathcal{F}$ must be estimable; that is, the design must have 100% estimation capacity. Thus, we focus on full estimation capacity designs and further distinguish them using the model-discrimination criteria of (21)–(24). As shown below, many full estimation capacity designs do not have good model discrimination capability. In practice, it is best if only full estimation capacity designs are used when model uncertainty is present. Whenever possible, only designs that are also model-discriminating designs should be used.

## 4.1 12-Run Model-Discriminating Designs

In Table 4, the values of the four measures (21)–(24) are reported for full estimation capacity, 12-run orthogonal designs with $f = 4, 5, 6, 7$ factors, respectively, and model space $\mathcal{F}_q$ of (8) with $q = 1, 2$. The designs are those described in Section 3.1 which consist of columns from the Hadamard matrix in Appendix A. When $q = 1$ two-factor interaction is to be estimated, all $f$-factor designs are full estimation capacity designs; when $q = 2$, the designs for which $f < 6$ satisfy $EC = 100\%$, as does the first design with $f = 6$ (see Section 3.1). For the second six-factor design and for the seven-factor design, both the min SA and min EPD are 0. When this happens, there exists at least one pair of models that are fully aliased and hence such designs are not recommended in practice.

TABLE 4. Measures of discriminating capability of 12-run designs

| | $q = 1$ | | | | $q = 2$ | | | |
|---|---|---|---|---|---|---|---|---|
| $f$ | min SA | ave SA | min EPD | ave EPD | min SA | ave SA | min EPD | ave EPD |
| 4 | 1.13 | 1.37 | .136 | .158 | 1.14 | 1.40 | .138 | .220 |
| 5 | 0.84 | 1.34 | .093 | .098 | 0.79 | 1.38 | .083 | .220 |
| 5 | 1.23 | 1.23 | .148 | .148 | 1.05 | 1.28 | .125 | .204 |
| 6 | 0.93 | 1.22 | .107 | .143 | 0.84 | 1.28 | .093 | .200 |
| 6 | 0.00 | 1.17 | .000 | .138 | — | — | — | — |
| 7 | 0.00 | 1.17 | .000 | .131 | — | — | — | — |

The designs for $f \geq 8$ all have fully aliased model pairs and are not included in Table 4.

It is interesting to note the difference between the two 5-factor designs. They both have $EC = 100\%$ and almost identical $IC$ values for each $q$ (see Sun, 1993). However, Table 4 shows that the second design has better model discrimination capabilities. In summary, when 12-run designs are used for screening purposes, they should be used for no more than six factors. To investigate five or six factors, the second 5-factor design and the first 6-factor design in Table 4 should be used. Complete catalogues of 12-run designs are available from the author's Web site (http://www.csom.umn.edu/∼wli).

## *4.2 16-Run Model-Discriminating Designs*

Table 5 summarizes the results for the 16-run orthogonal designs described in Section 2.3. From Li and Nachtsheim (2000), it can be verified that full estimation capacity designs exist for $f \leq 9$ when $q = 1$ and for $f \leq 8$ when $q = 2$. It is of interest to find the subset of designs that are model-discriminating. For each pair $(f, q)$ over the range $f = 5, \ldots, 9$ and $q = 1, 2$, Table 5 shows the number, $n_E$, of full estimation capacity designs, and the number, $n_M$, of these that are model-discriminating. For example, for $f = 5$ and $q = 1$, there are eight full estimation capacity designs, six of which are model-discriminating designs. The indices for model-discriminating designs are given in the fourth column in Table 5. For example, when $f = 5$ and $q = 2$, the five model-discriminating designs are Designs 4, 8, 10, 5, and 11. They are ranked by sequentially maximising (min SA, ave SA) of (21) and (22). The details of the resulting designs are provided in Appendix C. For example, under $f = 5$ in Appendix C, it is shown that Design 4 is obtained by using columns 1, 2, 4, 8, and 15 of Hadamard matrix H16.1 in Appendix B.

Several interesting issues arise from Table 5. First, the number of designs that are suitable for the purpose of model discrimination is very small compared with the

TABLE 5. Design indices for 16-run full estimation capacity designs that are model discriminating or $(EC, IC)$-optimal; $n_E$ and $n_M$ denote, respectively, the number of full estimation capacity designs and the number of these designs that are model discriminating

| $q$ | $f$ | $n_E$ | $n_M$ | Model-discriminating | $(EC, IC)$-optimal |
|---|---|---|---|---|---|
| 1 | 5 | 8 | 6 | 4,8,10,5,11,7 | 3,4 |
| | 6 | 16 | 8 | 26,24,27,13,20,19,22,23 | 5 |
| | 7 | 27 | 10 | 51,52,32,55,45,49,53,43,54,50 | 6 |
| | 8 | 16 | 3 | 67,68,72 | 6 |
| | 9 | 4 | 0 | — | 68 |
| 2 | 5 | 6 | 5 | 4,8,10,5,11 | 4 |
| | 6 | 8 | 2 | 26,27 | 13,19 |
| | 7 | 10 | 1 | 52 | 32,49 |
| | 8 | 3 | 0 | — | 67,68 |
| | 9 | 0 | — | — | — |

total number of candidate orthogonal designs. For $f = 5, 6, 7$, and 8, there are 11, 27, 55, and 80 orthogonal designs, respectively (see Table 1). Among these designs, there are only 6, 8, 10, and 3 model-discriminating designs, respectively, for $q = 1$. When $q = 2$, the numbers of designs are further reduced to 5, 2, 1, and 0, respectively. Secondly, nonregular designs (with no defining relation) are usually better than regular designs (with a defining relation) for the purpose of model discrimination. Among all model-discriminating designs displayed in Table 5, only two designs (Designs 4 and 5 for $f = 5$) are regular designs. (To find which design is regular, see Appendix C, where only projections of Hadamard matrix I result in regular designs.) In fact, when $f \geq 6$, all model-discriminating designs are nonregular designs. Thirdly, the $(EC, IC)$-optimal designs may not be model-discriminating designs. For example, when $f = 6$ and $q = 2$, the $(EC, IC)$-optimal designs are Designs 13 and 19. In contrast, the model-discriminating designs are Designs 26 and 27. Further investigation reveals that Designs 26 and 27 are not much worse than Designs 13 and 19 in terms of the $(EC, IC)$ criterion. The former has $(EC, IC)$ values of (1.00,0.88) and the latter has values (1.00,0.94). Some authors have advocated the use of $IC$ as the second criterion for distinguishing between designs having the same $EC$ value (for example, Li and Nachtsheim, 2000). This example demonstrates that the model discrimination criterion may be a better alternative in screening for model selection.

## *4.3 20-Run Model-Discriminating Designs*

Table 6 summarizes the model discrimination capabilities of the 20-run orthogonal designs discussed in Section 2.3. Table 3 shows that most 20-run designs have high $EC$ values. For example, for models with $q = 1$ two-factor interaction, $f$-factor designs have $EC = 100\%$ for all $f \geq 4$. However, the numbers of model-discriminating designs are much smaller. For example, for $f = 11, q = 1$, only 8.3% of the total designs are model-discriminating designs. When $f > 11$, none of the orthogonal designs are model-discriminating designs, even though they are all full estimation capacity designs. For each pair $(f, q)$, Table 6 also

TABLE 6. Summary of the best 20-run designs for model discrimination; $p_q$ is the percentage of model-discriminating designs among orthogonal designs

| | $q = 1$ | | | $q = 2$ | | |
|---|---|---|---|---|---|---|
| $f$ | $p_1(\%)$ | min SA | Index | $p_2(\%)$ | min SA | Index |
| 4 | 100 | 1.35 | 3 | 100 | 1.35 | 3 |
| 5 | 100 | 1.29 | 10,11 | 100 | 1.28 | 11 |
| 6 | 100 | 1.23 | 74,75 | 100 | 1.14 | 75 |
| 7 | 100 | .94 | 452 | 100 | .83 | 452 |
| 8 | 100 | .88 | 855 | 99.7 | .76 | 1370 |
| 9 | 97.1 | .93 | 2477 | — | — | — |
| 10 | 77.6 | .76 | 104 | — | — | — |
| 11 | 8.3 | .94 | 4,7,8,13 | — | — | — |

gives the design indices for the best model-discriminating designs, which are obtained by sequentially maximising (min SA, ave SA). For example, for $f = 6$ and $q = 1$, the best model-discriminating designs are Designs 74 and 75, both of which have min SA of 1.29 and the ave SA of 1.43. All 20-run orthogonal designs, including these model-discriminating designs, are available at the author's Web site.

The results for models with $q = 2$ two-factor interactions are similar to those for $q = 1$. Model-discriminating designs exist for $f \leq 11$. When producing the results for Table 6, I did not run the complete search for $f = 9$, 10, and 11, due to the extremely large amount of computing time that would be required. However, the results from the evaluation of randomly selected designs show that the percentages of model-discriminating designs for $q = 2$ are similar to those for $q = 1$.

## 5 Nonorthogonal Designs

In some situations, nonorthogonal designs may be better than orthogonal designs. For instance, it has been found that sometimes a small sacrifice of orthogonality can result in a design with greater capability for screening; see, for example, Li and Nachtsheim (2000) and Miller and Sitter (2005).

To find the best nonorthogonal designs, the exhaustive search method used in the previous two sections is usually not possible because the number of candidate designs may be too large. One commonly used methodology is the algorithmic approach, which is explained in Section 5.1 in the context of constructing model-robust factorial designs. Then, in Section 5.2, a nonalgorithmic approach using the foldover technique is discussed. In both sections, we focus on designs, in which each column of the design matrix has the same number of $+1$s and $-1$s. In Section 5.3, a Bayesian approach is introduced.

### *5.1 Optimal Designs Using Exchange Algorithms*

To search for an optimal design among balanced (mean orthogonal) designs for a given model space $\mathcal{F}$ and a criterion in $\mathcal{C}$, an algorithmic approach is usually appropriate. Most available algorithms aim to improve a design by changing either rows or columns of the design matrix. The columnwise-pairwise (CP) algorithms of Li and Wu (1997), which were proposed for the construction of optimal supersaturated designs, can retain the balance property during the optimization process for finding an optimal design. When the number of models in the model space $\mathcal{F}$ is large, the evaluation of the criterion in $\mathcal{C}$ may take a large amount of computing time. This motivated Li and Nachtsheim (2000) to propose a restricted CP algorithm, which can reduce the computing time substantially.

The restricted CP algorithm can be summarized as follows for two-level designs with coded levels $+1$ and $-1$ (Li and Nachtsheim, 2000, Appendix B).

Step 1: Choose a balanced (mean orthogonal) design at random; label the design matrix **D**.
Step 2: Set $j = 0$.
Step 3: Set $j = j + 1$. Randomly choose an element $d_{ij}$ from the $j$th column of **D**. Consider swaps of $d_{ij}$ and each other element in the same column. Find $i^*$ such that an exchange of $d_{ij}$ and $d_{i*j}$ results in a maximum gain in the criterion value.
Step 4: Repeat Step 3 until $j = f + 1$. If the new design matrix after $f$ exchanges makes negligible improvement, stop. Otherwise, repeat Steps 2–4.
Step 5: Record the optimal design. Repeat Steps 1–4 several times, the number of which is given in advance. Then, choose the best among those resulting designs.

This algorithm is also applicable to classes of designs where factors have more than two levels. In the case of two-level designs, the exchange procedure described in Step 3 amounts to swapping the sign of elements in a column. A MATLAB file is available from the Web site: http://www.csom.umn.edu/∼wli. Using the algorithm, Li and Nachtsheim (2000) constructed a class of model-robust factorial designs, which maximise the $(EC, IC)$ criterion among balanced (mean orthogonal) designs. Because the class of orthogonal designs is a subset of the class of balanced designs, the model-robust factorial design, chosen from balanced designs, is at least as good as the corresponding model-robust orthogonal design. In many cases, model-robust factorial designs have much larger $EC$ values than competing orthogonal designs and, sometimes, the difference can be substantial. Consider, for example, the 16-run design with nine factors. When $q = 3$ two-factor interactions are estimated, the model-robust orthogonal design has $EC_3 = 0.661$. In comparison, the model-robust factorial design has $EC_3 = 1.0$ (see Table 4 of Li and Nachtsheim, 2000).

Li and Nachtsheim (2000) compared the $(EC, IC)$ values for model-robust orthogonal designs and model-robust factorial designs with 12 runs and 16 runs, respectively. Their Tables 3 and 4 demonstrated that, in many cases, the full estimation capacity orthogonal designs do not exist, but the full estimation capacity balanced designs are available. For instance, for the 16-run designs with nine factors, the full estimation capacity orthogonal design does not exist for $q = 2$ or $q = 3$ but, in both cases, the model-robust factorial designs have $EC = 100\%$. Because it is important that full estimation capacity designs be used for model screening, the model-robust factorial designs can be very useful when no full estimation capacity orthogonal designs exist. To measure the model discrimination capabilities of the designs proposed in Li and Nachtsheim (2000), I computed the values of (21)–(24) for those designs and found that they generally perform quite well. For instance, consider the design for $n = 16$ and $f = 8$ shown in Figure 1 of Li and Nachtsheim (2000), which is presented here in Table 7. When $q = 2$, this design has min SA of .32 and min EPD of 0.20. It was demonstrated in Section 4.2. (see Table 5) that all 16-run full estimation capacity orthogonal designs with eight factors have min $SA = 0$. Thus, the 16-run with eight factors model-robust

TABLE 7. A 16-run model-robust factorial design with eight factors

| Factors | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| −1 | 1 | −1 | 1 | 1 | 1 | 1 | −1 |
| 1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 |
| 1 | 1 | 1 | 1 | −1 | 1 | 1 | 1 |
| 1 | −1 | 1 | −1 | −1 | −1 | −1 | −1 |
| −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 |
| 1 | −1 | −1 | 1 | 1 | −1 | 1 | 1 |
| −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 |
| −1 | −1 | −1 | −1 | 1 | −1 | −1 | −1 |
| −1 | −1 | 1 | 1 | −1 | −1 | −1 | 1 |
| 1 | −1 | 1 | 1 | −1 | 1 | −1 | −1 |
| −1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 |
| 1 | 1 | −1 | −1 | 1 | 1 | 1 | −1 |
| 1 | 1 | 1 | −1 | 1 | −1 | 1 | −1 |
| −1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 |
| 1 | −1 | −1 | −1 | −1 | 1 | −1 | 1 |
| −1 | 1 | 1 | −1 | 1 | 1 | −1 | 1 |

factorial design of Table 7 is superior to any of the corresponding orthogonal designs.

## 5.2 *A Non-algorithmic Approach Using Foldovers*

Many efficient designs can be produced by an algorithmic approach. However, there are many other types of designs that can be constructed through operations on the existing classes of designs. The literature is replete with such non-algorithmic approaches. For example, Lin (1993) and Wu (1993) both constructed efficient supersaturated designs from well-known Hadamard matrices. Such types of designs may not be optimal under a given criterion but they are easy to construct and they usually have some desirable properties. In the area of screening designs for model selection, one such approach was proposed by Miller and Sitter (2005), and this is described briefly below.

The approach of Miller and Sitter (2005) is focused on the use of foldover designs. For a design $d$ with design matrix $\boldsymbol{D}$, if the signs of some columns of $\boldsymbol{D}$ are reversed to obtain $\boldsymbol{D}^R$, then the resulting design composed of $\boldsymbol{D}$ and $\boldsymbol{D}^R$ is called a *foldover design*. For a review of foldover designs, see Chapter 1. Among related studies, Diamond (1995) investigated the projection properties of the foldover of the 12-run Plackett–Burman design. Miller and Sitter (2001) demonstrated that the "folded over" 12-run Plackett–Burman design is useful for considering main effects of up to 12 factors plus a few two-factor interactions. They considered only the full foldover, where the signs of all columns of $\boldsymbol{D}$ are reversed. For full foldover designs, the columns corresponding to two-factor interactions are not

orthogonal to each other, but they are all orthogonal to the columns corresponding to main effects. Thus, the identification of active main effects is not affected by the presence of active two-factor interactions. (An effect that is large enough to be of practical importance is called an *active* effect see Chapter 8). Miller and Sitter (2001) proposed a two-step analysis approach to take advantage of this property of the full foldover designs. Foldover designs can also be obtained by switching some, but not all, of the all columns in $\boldsymbol{D}$. The optimal foldover of regular and nonregular designs (in terms of the aberration criterion) were constructed by Li and Lin (2003), and Li et al. (2003), respectively. They found that many optimal foldovers are not full foldovers.

One limitation on the foldover of a Plackett–Burman design is that the number of runs of the combined design, resulting from the initial design plus its foldover, may be too large. Miller and Sitter (2005) extended the previous work of Miller and Sitter (2001) to the foldover of nonorthogonal designs. For example, they constructed a 12-run design by switching the signs of all columns of a $6 \times 5$ design. Miller and Sitter (2005) compared the folded over 12-run design with other competing designs, including the two 12-run five-factor orthogonal designs discussed in Section 2.3. Miller and Sitter (2001, 2005) demonstrated that the foldover design performs quite well in terms of the $EC$ criterion and the probability of identifying the correct model.

For nonorthogonal designs constructed by the full foldover, the columns corresponding to main effects are not orthogonal to each other. However, they are orthogonal to all the columns corresponding to two-factor interactions. Consequently, the design may actually have a higher ability to identify active main effects than competing orthogonal designs while maintaining the ability to identify a few active two-factor interactions. For more details, see Miller and Sitter (2005).

## 5.3 A Bayesian Modification of the D-Optimal Approach

DuMouchel and Jones (1994) proposed a Bayesian modification of the search for $D$-optimal designs in order to address model uncertainties. Consider the usual model of (2), $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is an $n \times h$ model matrix. Suppose, in addition to the $h$ parameters in the model, which were called *primary* terms by DuMouchel and Jones (1994), there are $q$ potential terms that are just possibly important. In their work DuMouchel and Jones assumed that $\sigma^2 = 1$, coefficients of primary terms have a diffuse prior (that is, the prior variance tends to infinity) with an arbitrary prior mean, and coefficients of potential terms are independent, have a prior mean of 0, and a finite variance $\tau^2$. Let $\mathbf{K}$ be the $(p+q) \times (p+q)$ diagonal matrix whose first $p$ diagonal elements are equal to 0 and whose last $q$ diagonal elements are 1. Under the model assumptions, $\epsilon \sim N(o, \sigma^2 I)$ and $\boldsymbol{\beta} \sim N(0, \tau^2 \boldsymbol{K})$, $\boldsymbol{\beta}|\boldsymbol{y} \sim N(\mathbf{b}, (\mathbf{X'X} + \mathbf{K}/\tau^2)^{-1})$, where $\mathbf{b} = (\mathbf{X'X} + \mathbf{K}/\tau^2)^{-1}\mathbf{X'y}$, and thus a Bayes $D$-optimal design would maximise $|\mathbf{X'X} + \mathbf{K}/\tau^2|$ which leads to the selection of different designs (see DuMouchel and Jones, 1994). This approach can preserve

the flexibility and ease of use of algorithmic designs while being more resistant to the bias caused by an incorrect model.

## 6 Discussion

In this chapter, some useful results and methods for the construction of screening designs for model selection have been described. A general framework for finding designs has been provided which consists of a model space, a criterion for design choice, and a set of candidate designs. Several useful criteria and model spaces for model selection have been considered, with the emphasis on finding designs that maximise estimation capacity and model discrimination capability. Both efficient orthogonal and nonorthogonal designs have been investigated. For orthogonal designs, new results have been given, including the tabulation of effective 12-, 16-, and 20-run designs for practical use. I recommend the use of model-discriminating designs for screening experiments. Most of the results in this chapter have been obtained using a model space consisting of models with all main effects and a small number of two-factor interactions. However, the methods described here are applicable to other model spaces as well.

Although this chapter is focused on model estimation and model discrimination criteria, other criteria may also be useful. For example, Lewis and Dean (2001) proposed a new strategy for using group screening to estimate both main effects and key two-factor interactions. Their criteria minimise the expected total number of observations, the probability that the size of the experiment exceeds a pre-specified target, and the proportion of active effects that are not detected (see also Chapter 9).

## Appendix A
## 12-Run Hadamard Matrix

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | −1 | 1 | 1 | 1 | −1 | −1 | −1 | 1 | −1 |
| 1 | −1 | 1 | 1 | −1 | 1 | 1 | 1 | −1 | −1 | −1 | 1 |
| 1 | 1 | −1 | 1 | 1 | −1 | 1 | 1 | 1 | −1 | −1 | −1 |
| 1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | 1 | 1 | −1 | −1 |
| 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | 1 | 1 | −1 |
| 1 | −1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | 1 | 1 |
| 1 | 1 | −1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | 1 |
| 1 | 1 | 1 | −1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 |
| 1 | 1 | 1 | 1 | −1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 |
| 1 | −1 | 1 | 1 | 1 | −1 | −1 | −1 | 1 | −1 | 1 | 1 |
| 1 | 1 | −1 | 1 | 1 | 1 | −1 | −1 | −1 | 1 | −1 | 1 |
| 1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 |

# Appendix B
# Hall's 16-Run Orthogonal Designs

**H16.I: 1,2,4,8 are independent columns**

$$\mathbf{1},\mathbf{2},3 = 12,\mathbf{4},5 = 14,6 = 24,7 = 124,\mathbf{8},$$
$$9 = 18,10 = 28,11 = 128,12 = 48,13 = 148,14 = 248,15 = 1248$$

**H16.II:**

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 |
| 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 |
| 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 |
| 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 |
| 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 |
| 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 |
| 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 |
| 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 |
| 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 |
| 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 |
| 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 |
| 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 |
| 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 |
| 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 |
| 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 |

**H16.III:**

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 |
| 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 |
| 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 |
| 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 |
| 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 |
| 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 |
| 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 |
| 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 |
| 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 |
| 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 |
| 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 |
| 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 |
| 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 |
| 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 |
| 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 |

**H16.IV:**

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 |
| 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 |
| 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 |
| 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 |
| 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 |
| 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 |
| 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 |
| 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 |
| 1 | −1 | 1 | −1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 |
| 1 | −1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | −1 | 1 |
| 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 |
| 1 | −1 | −1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 | 1 | −1 |
| 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 |
| 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 |
| 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | 1 | −1 |

**H16.V:**

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 |
| 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 |
| 1 | 1 | 1 | 1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 |
| 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 |
| 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 |
| 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 |
| 1 | 1 | −1 | −1 | −1 | −1 | 1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 |
| 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | 1 | 1 | 1 | −1 | −1 | −1 | 1 | 1 |
| 1 | −1 | 1 | −1 | 1 | −1 | 1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | −1 | −1 |
| 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 |
| 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 |
| 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 | −1 | 1 |
| 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 |
| 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 |
| 1 | −1 | −1 | 1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 | −1 | 1 | 1 | −1 |

# Appendix C
# Selected 16-Run Orthogonal Designs (Expressed as Projections of $H_{16.I} - H_{16.V}$)

$f = 5$:

| | | |
|---|---|---|
| 4 = I(1,2,4,8,15); | 5 = II(1,2,4,8,12); | 7 = II(1,4,6,8,12); |
| 8 = II(4,5,6,8,12); | 10 = II(4,5,8,10,12); | 11 = III(2,4,8,10,12); |

$f = 6$:

| | | |
|---|---|---|
| 3 = I(1,2,3,4,8,12); | 5 = I(1,2,4,7,8,11); | 13 = II(1,4,6,8,11,12); |
| 19 = III(1,2,4,8,10,12); | 22 = III(2,4,7,8,10,12); | 23 = III(2,4,8,9,10,12); |
| 24 = III(2,4,8,9,10,14); | 26 = III(2,4,8,10,12,15); | 27 = IV(2,4,6,8,10,12); |

$f = 7$:

| | | |
|---|---|---|
| 6 = I(1,2,4,7,8,11,13); | 32 = III(1,2,4,8,10,12,15); | 43 = III(2,4,7,8,9,10,12); |
| 45 = III(2,4,7,8,10,12,15); | 49 = IV(1,2,4,6,8,10,12); | 50 = IV(2,3,4,6,8,10,12); |
| 51 = IV(2,4,6,8,10,12,14); | 52 = IV(2,4,6,8,10,12,15); | 53 = V(1,2,4,8,9,10,12); |
| 54 = V(1,2,4,8,9,10,13); | 55 = V(1,2,4,8,10,12,15); | |

$f = 8$:

| | | |
|---|---|---|
| 6 = I(1,2,4,7,8,11,13,14); | 67 = IV(1,2,4,6,8,10,12,14); | 68 = IV(1,2,4,6,8,10,12,15); |
| 72 = IV(2,3,4,6,8,10,12,14); | | |

$f = 9$:

| | | |
|---|---|---|
| 4 = I(1,2,3,4,5,8,9,14,15); | 71 = IV(1,2,3,4,6,8,10,12,14); | |
| 73 = IV(2,3,4,5,6,7,8,12,14); | 76 = IV(2,3,4,5,6,8,10,12,14); | |
| 85 = V(1,2,4,8,9,10,11,12,13) | | |

$f = 10$:

| | |
|---|---|
| 66 = IV(2,3,4,5,6,7,8,10,12,14); | 76 = V(1,2,4,7,8,9,10,11,12,13); |

## *References*


Addelman, S. (1962). Symmetrical and asymmetrical fractional factorial plans. *Technometrics*, **4**, 47–58.

Allen, T. T. and Bernshteyn, M. (2003). Supersaturated designs that maximize the probability of identifying active factors. *Technometrics*, **45**, 92–97.

Bingham, D. and Li, W. (2002). A class of optimal robust parameter designs. *Journal of Quality Technology*, **34**, 244–259.

Biswas, A. and Chaudhuri, P. (2002). An efficient design for model discrimination and parameter estimation in linear models. *Biometrika*, **89**, 709–718.

Cheng, C. S., Steinberg, D. M., and Sun, D. X. (1999). Minimum aberration and model robustness for two-level fractional factorial designs. *Journal of the Royal Statistical Society B*, **61**, 85–93.

Diamond, N. T. (1995). Some properties of a foldover design. *Australian Journal of Statistics*, **37**, 345–352.

DuMouchel, W. and Jones, B. (1994). A simple Bayesian modification of *D*-optimal designs to reduce dependence on an assumed model. *Technometrics*, **36**, 37–47.

Hall, M. J. (1967). *Combinatorial Theory*. Blaisdell, Waltham, MA.

Jones, B., Li, W., Nachtsheim, C. J., and Ye, K. Q. (2005). Model discrimination—Another perspective of model-robust designs. *Journal of Statistical Planning and Inferences*. In press.

Läuter, E. (1974). Experimental design in a class of models. *Mathematische Operationsforschung und Statistik*, **5**, 379–398.

Lewis, S. M. and Dean, A. M. (2001). Detection of interactions in experiments on large numbers of factors. *Journal of the Royal Statistical Society B*, **63**, 633–672.

Li, W. and Lin, D. K. J. (2003). Optimal foldover plans for two-level fractional fractorial designs. *Technometrics*, **45**, 142–149.

Li, W. and Nachtsheim, C. J. (2000). Model-robust factorial designs. *Technometrics*, **42**, 345–352.

Li, W. and Nachtsheim, C. J. (2001). Model-robust supersaturated and partially supersaturated designs. *Technical Report* UMSI 2001/3. University of Minnesota Supercomputing Institute, Minneapolis, MN 55455.

Li, W. and Wu, C. F. J. (1997). Columnwise-pairwise algorithms with applications to the construction of supersaturated designs. *Technometrics*, **39**, 171–179.

Li, W., Lin, D. K. J., and Ye, K. Q. (2003). Optimal foldover plans for nonregular designs. *Technometrics*, **45**, 347–351.

Lin, D. K. J. (1993). A new class of supersaturated designs. *Technometrics*, **35**, 28–31.

Lin, D. K. J. and Draper, N. R. (1992). Projection properties of Plackett and Burmann designs. *Technometrics*, **34**, 423–428.

Liu, Y. F. and Dean, A. M. (2004). k-Circulant supersaturated designs. *Technometrics*, **46**, 32–43.

MATLAB (2004): Release 7.01, The Math Works, Inc.

Miller, A. and Sitter, R. R. (2001). Using the folded-over 12-run Plackett-Burman designs to consider interactions. *Technometrics*, **43**, 44–55.

Miller, A. and Sitter, R. R. (2005). Using folded-over nonorthogonal designs. *Technometrics*. **47**, 502–513.

Myers, R. H. and Montgomery, D. C. (2002). *Response Surface Methodology: Process and Product Optimisation Using Designed Experiments*. John Wiley and Sons, New York.

Srivastava, J. N. (1975). Designs for searching non-negligible effects. In *A Survey of Statistical Designs and Linear Models*. Editor: J. N. Srivastava. North-Holland, Amsterdam.

Sun, D. X. (1993). Estimation capacity and related topics in experimental designs. Ph.D. Thesis, University of Waterloo, Ontario, Canada.

Sun, D. X., Li, W., and Ye, Q. K. (2002). An algorithm for sequentially constructing non-isomorphic orthogonal designs and its applications. *Technical Report* SUNYSB-AMS-02-13, Department of Applied Mathematics and Statistics, SUNY at Stony Brook, New York.

Taguchi, G. (1987). *System of Experimental Desion*. Two volumes. Unipub/Kraus International, White Plains, N.

Wu, C. F. J. (1993). Construction of supersaturated designs through partially aliased interactions. *Biometrika*, **80**, 661–669.

Wu, C. F. J, and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. Wiley, New York.

# 11
# Prior Distributions for Bayesian Analysis of Screening Experiments

HUGH CHIPMAN

When many effects are under consideration in a screening experiment, it may be necessary to use designs with complex aliasing patterns, especially when interactions and higher-order effects exist. In this situation, the selection of subsets of active effects is a challenging problem. This chapter describes Bayesian methods for subset selection, with emphasis on the choice of prior distributions and the impact of this choice on subset selection, computation, and practical analysis. Attention is focused on experiments where a linear regression model with Gaussian errors describes the response. Ideas are illustrated through an experiment in clinical laboratory testing and through an example with simulated data. Advantages of the Bayesian approach are stressed, such as the ability to incorporate useful information about which subsets of effects are likely to be active. For example, an $AB$ interaction effect might only be considered active if main effects for $A$ and $B$ are also likely to be active. When such information is combined with a stochastic search for promising subsets of active effects, a powerful subset selection tool results. The techniques may also be applied to designs without complex aliasing as a way of quantifying uncertainty in subset selection.

## 1 Introduction

Many of the ideas in this chapter are motivated through the discussion, below, of an example of a screening experiment. This discussion is then followed by an overview of the rest of the chapter.

### *1.1 A Blood-Glucose Screening Experiment*

An experiment to study the effect of eight factors on blood-glucose readings made by a clinical laboratory testing device was described by Henkin (1986). The factor descriptions and levels are given in Table 1. One factor, $A$, has two levels whereas each of the other seven factors, $B - H$, has three levels. The design of the experiment had 18 runs and is shown, together with the data, in Table 2.

A goal of such screening experiments is the identification of the active effects. The concept of the activity of an effect was introduced by Box and Meyer (1986) who presented one of the first Bayesian methods for the analysis of designed experiments; see also Chapter 8. The general approach described in this chapter

TABLE 1. Factors for the glucose experiment

| Factor | Description | Levels |
|---|---|---|
| *A* | Wash | yes, no |
| *B* | Volume in microvial | 2.0, 2.5, 3.0 ml |
| *C* | Water level in caras | 20.0, 28.0, 35.0 ml |
| *D* | Speed of centrifuge | 2100, 2300, 2500 RPM |
| *E* | Time in centrifuge | 1.75, 3.00, 4.50 minutes |
| *F* | (Sensitivity, absorption) | (0.10,2.5), (0.25,2.0), (0.50,1.5) |
| *G* | Temperature | 25, 30, 37°*C* |
| *H* | Dilution | 1:51, 1:101, 1:151 |

is to view the analysis of screening data as a regression problem in which an $n \times (h+1)$ matrix, $\boldsymbol{X}$, of predictors or explanatory variables is constructed with the last $h$ columns being contrasts in the levels of the factors (see Chapter 1). For example, for the glucose data, linear and quadratic main effects and interaction effects are considered and these effects are represented by a vector of regression coefficients $\boldsymbol{\beta}$. The first column of $\boldsymbol{X}$ is a vector of 1s and correspondingly, the first element of $\boldsymbol{\beta}$ is an intercept. Most of the effects are assumed to be *inactive* (near zero). The task of identifying a subset of active effects corresponds to identifying which contrasts in $\boldsymbol{X}$ should be included in a regression model.

The selection of factorial effects may seem counterintuitive for screening experiments, because the primary goal is to identify important factors. The philosophy behind the identification of individual active effects is that, once active effects are

TABLE 2. The design and response data for the glucose experiment

| Factor | | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|
| A | G | B | C | D | E | F | H | reading |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 97.94 |
| 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 83.40 |
| 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 95.88 |
| 1 | 3 | 1 | 1 | 2 | 2 | 3 | 3 | 88.86 |
| 1 | 3 | 2 | 2 | 3 | 3 | 1 | 1 | 106.58 |
| 1 | 3 | 3 | 3 | 1 | 1 | 2 | 2 | 89.57 |
| 1 | 2 | 1 | 2 | 1 | 3 | 2 | 3 | 91.98 |
| 1 | 2 | 2 | 3 | 2 | 1 | 3 | 1 | 98.41 |
| 1 | 2 | 3 | 1 | 3 | 2 | 1 | 2 | 87.56 |
| 2 | 1 | 1 | 3 | 3 | 2 | 2 | 1 | 88.11 |
| 2 | 1 | 2 | 1 | 1 | 3 | 3 | 2 | 83.81 |
| 2 | 1 | 3 | 2 | 2 | 1 | 1 | 3 | 98.27 |
| 2 | 3 | 1 | 2 | 3 | 1 | 3 | 2 | 115.52 |
| 2 | 3 | 2 | 3 | 1 | 2 | 1 | 3 | 94.89 |
| 2 | 3 | 3 | 1 | 2 | 3 | 2 | 1 | 94.70 |
| 2 | 2 | 1 | 3 | 2 | 3 | 1 | 2 | 121.62 |
| 2 | 2 | 2 | 1 | 3 | 1 | 2 | 3 | 93.86 |
| 2 | 2 | 3 | 2 | 1 | 2 | 3 | 1 | 96.10 |

identified, the important factors involved in these effects will also be known. An emphasis on the selection of active effects instead of active factors means that more insight is gained into the nature of the relationship between the factors and the response.

In the glucose experiment, factor $A$ has two qualitative levels and so there is only one contrast with $+1$ and $-1$ corresponding to the levels "yes" and "no". All the three-level factors are quantitative and their main effects can be decomposed into linear and quadratic effects using orthogonal polynomials; see, for example, Draper and Smith (1998, Chapter 22) and Dean and Voss (1999, page 71). For factors $B$, $D$, $F$, and $H$, with evenly spaced levels, the respective coefficients for the low, middle, and high factor levels in each linear contrast are $(-1, 0, 1)/\sqrt{2}$ and in each quadratic contrast are $(1, -2, 1)/\sqrt{6}$. Factor $F$ combines two variables (sensitivity and absorption), and has slightly nonuniform spacing of sensitivity. Because a single natural scale is difficult to specify for such a combined factor, the contrasts used here to measure its main effect are identical to those used for a single evenly spaced factor. For the unevenly spaced factors $C$, $E$, and $G$, linear contrast coefficients depend on the spacing of the factor levels and are calculated as the original values minus their means. This is called "centering". Quadratic contrast coefficients are formed by squaring the elements of the centered linear contrasts and then centering the squared values. So, for factor $C$ with levels $\{20, 28, 35\}$, the linear contrast $C_L$ has coefficients $(-7.67, 0.33, 7.33)$ and the quadratic contrast $C_Q$ has coefficients $(21.22, -37.44, 16.22)$.

The nonregular nature of the fractional factorial design makes it possible to consider interaction effects as well as main effects; see also Chapter 7. An interaction between two factors, each with three levels, has four degrees of freedom which can be decomposed into linear $\times$ linear, linear $\times$ quadratic, quadratic $\times$ quadratic, and quadratic $\times$ linear effects. The contrast coefficients for these effects are formed by multiplying the coefficients of the corresponding main effect contrasts.

In the glucose experiment, a total of $h = 113$ effects is under consideration. This includes 8 linear main effects $(A_L, \ldots, H_L)$, 7 quadratic main effects $(B_Q, \ldots, H_Q)$, 28 linear $\times$ linear interactions $(A_L B_L, \ldots, G_L H_L)$, $7 + 7 \times 6 = 49$ linear $\times$ quadratic interactions $(A_L B_Q, \ldots, A_L H_Q, B_L C_Q, \ldots, G_L H_Q)$ and 21 quadratic $\times$ quadratic interactions $(B_Q C_Q, \ldots, G_Q H_Q)$.

The contrasts described above are not scaled to be directly comparable. Because active effects are identified via regression modeling, detection of activity will be unaffected by scaling of contrasts. By taking a regression approach to the analysis of the glucose data in Table 2, the screening problem reduces to one of selecting a small subset of active effects from the 113 effects under consideration. Hamada and Wu (1992) tackled the subset selection problem for screening experiments using a modified stepwise regression procedure. They first identified active main effects and then identified active interactions between those active main effects. In the glucose experiment, the subset of active effects that they identified was $E_Q$, $F_Q$, and interaction $E_L F_L$. The model composed of these effects plus an overall mean has an $R^2$ value of 68%.

TABLE 3. Best fitting subsets of size 1–6, identified by "all subsets" search for the glucose experiment

| Subset of active effects | $R^2$ |
|---|---|
| $B_L H_Q$ | 46.2 |
| $B_L H_Q, B_Q H_Q$ | 77.0 |
| $B_L, B_L H_Q, B_Q H_Q$ | 85.5 |
| $E_Q, A_L C_L, B_L H_Q, B_Q H_Q$ | 94.3 |
| $F_L, A_L H_Q, G_L H_Q, B_L H_Q, E_L F_L$ | 97.0 |
| $A_L, E_Q, A_L C_L, B_L D_L, B_L H_Q, B_Q H_Q$ | 98.7 |

Although stepwise selection algorithms run quickly, they do not consider all possible subsets of active effects. Instead, they build up a model one term at a time. An additional problem with the stepwise approach of Hamada and Wu (1992) is that, by dividing the algorithm into stages, the search is further restricted: a highly significant (active) interaction with no corresponding main effects that are active will not be identified.

A remedy to the limited scope of a stepwise search is to use *all subsets regression*, in which regression models using every possible subset of active effects are considered. Furnival and Wilson (1974) developed an efficient algorithm for this computation. For the glucose data, a search over all subsets with six or fewer active effects was carried out, using the `leaps` package in R (R Development Core Team, 2004). The computation took about 50 minutes on a 1 GHz Pentium-III computer. For each size, the subset with highest $R^2$ is shown in Table 3. The three-term model with effects $B_L$, $B_L H_Q$, and $B_Q H_Q$ has $R^2 = 85.5\%$ compared with $R^2 = 68\%$ for the Hamada–Wu model with $E_Q$, $F_Q$, and $E_L F_L$.

A problem with the "all subsets" approach is that all relationships between predictors are ignored. For example, the best subset of size four, ($E_Q$, $A_L C_L$, $B_L H_Q$, $B_Q H_Q$), contains an interaction involving factors $A$ and $C$, but no corresponding main effects. Indeed, one of the main strengths of the Hamada–Wu approach is the incorporation of the principle of *effect heredity*: an interaction between two effects is not considered active unless at least one of the corresponding main effects is also active.

Neither all subsets regression nor the Hamada–Wu stepwise algorithm represents a complete solution. All subsets regression provides a complete search but ignores effect heredity. The Hamada–Wu approach identifies models obeying effect heredity but has an incomplete search and may miss the best effect heredity models.

The use of Bayesian priors, coupled with efficient stochastic search algorithms, provides one approach that solves both problems. The stochastic search significantly improves the chances of finding good models, whereas Bayesian priors focus the search on models obeying effect heredity.

To give a flavor of the Bayesian approach, two summaries of the Bayesian analysis of the glucose data are now presented. A more detailed analysis, including a discussion of prior distributions and computational methods, is given in Section 5.2.

TABLE 4. Ten subsets of effects with largest posterior probability for the glucose experiment

| Subset | Posterior probability | $R^2$ |
|---|---|---|
| $B_L, B_L H_L, B_L H_Q, B_Q H_Q$ | 0.126 | 86.0 |
| $B_L H_Q, B_Q H_Q$ | 0.069 | 77.0 |
| $B_L, B_L H_Q, B_Q H_Q$ | 0.064 | 85.5 |
| $B_L, B_L H_L, B_Q H_L, B_L H_Q, B_Q H_Q$ | 0.037 | 88.3 |
| $F_L, H_L, H_Q, A_L H_Q, G_L H_Q, B_L H_Q, E_L F_L$ | 0.037 | 97.0 |
| $A_L, B_L, A_L D_Q, B_L H_L, B_L H_Q, B_Q H_Q$ | 0.020 | 95.0 |
| $F_L, H_L, H_Q, A_L H_Q, B_L H_Q, B_Q H_Q$ | 0.019 | 93.0 |
| $F_L, H_L, H_Q, A_L H_Q, B_L H_Q, B_Q H_Q, C_L H_Q$ | 0.019 | 95.9 |
| $B_L, B_Q, B_L H_L, B_L H_Q, B_Q H_Q$ | 0.018 | 89.2 |
| $F_L, H_L, H_Q, A_L H_Q, G_L H_Q, B_L H_Q, B_Q H_Q, E_L F_L$ | 0.017 | 97.8 |

Table 4 lists the 10 most probable subsets found by the Bayesian procedure. With the exception of the second subset listed ($B_L H_Q, B_Q H_Q$), every term in every subset has at least one lower-order effect also in the subset. For example, in the fifth subset listed in Table 4 ($F_L, H_L, H_Q, A_L H_Q, G_L H_Q, B_L H_Q, E_L F_L$), the active effect $G_L H_Q$ has "parent" $H_Q$ which, in turn, has parent $H_L$. (The notions of parents and effect heredity are stated precisely in Section 2.2.) This fifth subset contains all the effects in the best subset of size 5 listed in Table 3. The Bayesian procedure has found a subset similar to one of the best subsets but which obeys effect heredity.

The Bayesian approach is more than a tool for adjusting the results of the all subsets regression by adding appropriate effects to achieve effect heredity. Take, for example, the sixth model in Table 4 which consists of $A_L, B_L, A_L D_Q, B_L H_L, B_L H_Q, B_Q H_Q$. The $A_L D_Q$ effect identified as part of this model does not appear in the best subsets of size 1–6 in Table 3. The Bayesian procedure has therefore discovered an additional possible subset of effects that describes the data.

Figure 1 displays a second summary in the form of the marginal posterior probability that each of the 113 effects is active. Effects are ordered by this probability along the horizontal axis. The height of the vertical line for each effect represents the posterior probability of activity under a certain choice of prior hyperparameters (see Section 2.2). Robustness of these probabilities is indicated by the rectangles, which give minimum and maximum marginal posterior probability of activity over 18 different choices of prior hyperparameters. From this plot, it evident that effects $B_L H_Q$ and $B_Q H_Q$ are very likely to be active, as well as other effects involving $B_L$, $H_L$, and $H_Q$. In addition, effects $A_L H_Q$ and $F_L$ might be active.

## *1.2 Overview of the Chapter*

The Bayesian approach described in Section 1.1 can be applied to a wide variety of screening problems, including those with both quantitative and qualitative factors.

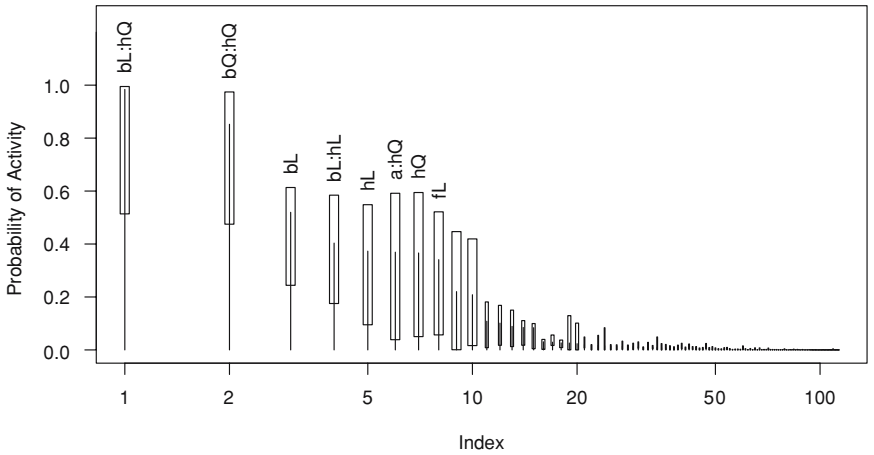



FIGURE 1. Marginal posterior probability of activity of the 113 effects for the glucose experiment; rectangles indicate robustness of the probabilities.

Although a large number of high-order contrasts was examined in the glucose example, this technique will work equally well when only linear main effects and linear $\times$ linear interactions are considered. When there are complex aliasing patterns between effects, the Bayesian approach is effective in identifying different subsets of effects that fit the data well. It can also be used in conjunction with regular fractional factorial designs.

Bayesian methods for subset selection offer several advantages over other approaches: the assignment of posterior probabilities to different subsets of active effects provides a way of characterizing uncertainty about effect activity; prior distributions can incorporate principles of effect dependence, such as effect heredity; the identification of promising models via Bayesian stochastic search techniques is faster than all subsets searches, and more comprehensive than stepwise methods.

The motivating example has illustrated the challenges of subset selection for screening experiments and the results of a Bayesian analysis. The remainder of the chapter provides details of the Bayesian approach, but some familiarity with Bayesian methods is assumed. Lee (2004) provides a good introduction to these methods without assuming too much advanced statistical background. Chapter 3 of Zellner (1987) provides detailed background on Bayesian multiple regression modeling. Bayesian simulation techniques, namely, Markov chain Monte Carlo (MCMC), are overviewed in Chapter 11 of Gelman et al. (2003) and in Chapter 10 of O'Hagan and Forster (2004).

Central to Bayesian approaches is the treatment of model parameters, such as the vector of regression coefficients $\boldsymbol{\beta}$, as random variables. Uncertainty and expert knowledge about these parameters are expressed via a prior distribution. The observed data give rise to a likelihood for the parameters. The likelihood and

prior distribution are combined to give a posterior distribution for the parameters. In subset selection for linear regression, the model is extended to include not only the regression coefficient vector $\boldsymbol{\beta}$ and the residual standard deviation $\sigma$, but also a vector $\boldsymbol{\delta}$ of binary indicators specifying whether each effect is active or inactive; that is, $\boldsymbol{\delta}$ identifies a subset of active effects. Interest focuses on prior and posterior distributions for $\boldsymbol{\beta}$, $\sigma$, and $\boldsymbol{\delta}$.

The Bayesian approach to subset selection is outlined in Sections 2 to 4. Section 2 gives the mathematical ingredients of the analysis: a probability model for the data, prior distributions for the parameters ($\boldsymbol{\beta}$, $\sigma$, $\boldsymbol{\delta}$) of the model, and the resultant posterior distribution.

For a particular data set, subsets with high posterior probability must be identified. This can be a computational challenge: with $h$ possible effects, there are $2^h$ different subsets. In order to identify promising subsets, MCMC methods for simulating from the posterior distribution on subsets may be used as a stochastic search. Section 3 outlines efficient techniques for exploring the subset space using MCMC methods.

Section 4 reviews simple, semi-automatic methods of choosing the hyperparameters of a prior distribution and adds some new insights into the choice of hyperparameters for a prior on regression coefficient vector $\boldsymbol{\beta}$. The glucose experiment and a simulated data set are used in Section 5 to demonstrate the application of the Bayesian subset selection technique.

A promising new use of prior distributions for subset selection is in the formulation of optimality criteria for the construction of designs that allow model discrimination. This technique is discussed in Section 6. The chapter concludes with a discussion, including possible extensions of the techniques to generalized linear models.

## 2 Model Formulation

This section gives a review of the linear regression model, including an augmented form that allows specification of the subset of active effects. Prior distributions are introduced and relevant posterior distributions given.

### *2.1 The Linear Regression Model*

The usual regression model,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon \tag{1}$$

is used, where $\boldsymbol{Y}$ is a vector of $n$ responses, $\boldsymbol{X}$ an $n \times (h+1)$ matrix of predictors, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_h)'$ a vector containing an intercept $\beta_0$ and $h$ regression coefficients, and $\varepsilon$ a vector of $n$ error variables $\varepsilon_1, \ldots, \varepsilon_n$ which are assumed to be independent and identically distributed as $N(0, \sigma^2)$. The elements of the columns of $\boldsymbol{X}$ are functions of the levels of one or more of the original factors and

may represent linear, quadratic, or interaction contrast coefficients, or indicator variables for qualitative factors.

This model is augmented by an unobserved indicator vector $\boldsymbol{\delta}$. Each element $\delta_j (j = 1, \ldots, h)$ of $\boldsymbol{\delta}$ takes the value 0 or 1, indicating whether the corresponding $\beta_j$ belongs to an inactive or an active effect, respectively. Because the intercept $\beta_0$ is always present in the model, it has no corresponding $\delta_0$ element. An inactive effect has $\beta_j$ close to 0 and an active effect has $\beta_j$ far from 0. The precise definition of "active" and "inactive" may vary according to the form of prior distribution specified. Under this formulation, the Bayesian subset selection problem becomes one of identifying a posterior distribution on $\boldsymbol{\delta}$.

## 2.2 *Prior Distributions for Subset Selection in Regression*

A Bayesian analysis proceeds by placing prior distributions on the regression coefficient vector $\boldsymbol{\beta}$, error standard deviation $\sigma$, and subset indicator vector $\boldsymbol{\delta}$. One form of prior distribution is given in detail below and other approaches are then discussed. Techniques for choosing hyperparameters of prior distributions, such as the mean of a prior distribution, are discussed later in Section 4.

A variety of prior distributions has been proposed for the subset selection problem and most differ in terms of the distributions placed on $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. One particular formulation, due to Box and Meyer (1986) and George and McCulloch (1997), is reviewed in detail here. Subsequent examples use this particular formulation, although many issues that arise are general.

The joint prior density on $\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}$ can be factored and subsequently simplified by assuming that the subset indicator vector $\boldsymbol{\delta}$ and error variance are independent. Then,

$$p(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}) = p(\boldsymbol{\beta}|\sigma, \boldsymbol{\delta})p(\sigma, \boldsymbol{\delta}) = p(\boldsymbol{\beta}|\sigma, \boldsymbol{\delta})p(\sigma)p(\boldsymbol{\delta}). \tag{2}$$

The prior densities $p(\boldsymbol{\beta}|\sigma, \boldsymbol{\delta})$, $p(\sigma)$, and $p(\boldsymbol{\delta})$ are described below.

### 2.2.1 Prior Distribution on Regression Parameters $\boldsymbol{\beta}, \sigma$

Prior distributions are often chosen to simplify the form of the posterior distribution. The posterior density is proportional to the product of the likelihood and the prior density and so, if the prior density is chosen to have the same form as the likelihood, simplification occurs. Such a choice is referred to as the use of a *conjugate* prior distribution; see Lee (2004) for details. In the regression model (1), the likelihood for $\boldsymbol{\beta}, \sigma$ can be written in terms of the product of a normal density on $\boldsymbol{\beta}$ and an inverse gamma density on $\sigma$. This form motivates the conjugate choice of a normal-inverse-gamma prior distribution on $\boldsymbol{\beta}, \sigma$. Additional details on this prior distribution are given by Zellner (1987).

The prior distribution used for the error variance is

$$\sigma^2 \sim \text{Inverse Gamma}(\nu/2, \nu\lambda/2) \tag{3}$$

and this is equivalent to specifying $\nu\lambda/\sigma^2 \sim \chi^2_\nu$. This prior distribution is identical to the likelihood for $\sigma^2$ arising from a data set with $\nu$ observations and sample variance $\lambda$. Although the data are normally distributed, the likelihood, when viewed as a function of $\sigma^2$, has this form.

A variety of prior distributions for $\boldsymbol{\beta}$ has been proposed in the literature. Here, the following formulation is used: for any given subset indicator vector $\boldsymbol{\delta}$ and value of the error variance $\sigma^2$, the effects $\beta_1, \ldots, \beta_h$ have independent normal prior distributions. The variance of $\beta_j (j = 1, \ldots, h)$ is formulated to be large or small depending on whether an effect is active ($\delta_j = 1$) or inactive ($\delta_j = 0$), through use of the hyperparameters $c_j$, $\tau_j$ in the distribution with density

$$p(\beta_j|\delta_j, \sigma) \sim \begin{cases} N(0, (\tau_j\sigma)^2) & \text{if } \delta_j = 0, \\ N(0, (c_j\tau_j\sigma)^2) & \text{if } \delta_j = 1. \end{cases} \tag{4}$$

This distribution is referred to as a "mixture of two normal distributions". The values of the hyperparameters $c_j$, $\tau_j$ are chosen to indicate magnitudes of inactive and active effects. Roughly speaking, $\beta_j$ for an active effect will be $c_j$ times larger than that for an inactive effect. Choosing $c_j$ much larger than 1 represents this belief. In Section 4, an automatic method of selecting values of the hyperparameters $c_j$ and $\tau_j$ is suggested. The intercept $\beta_0$ will have an uninformative prior distribution (i.e., $c_0 = 1$, $\tau_0 \to \infty$). Implicit in (4) is the assumption of a diagonal prior covariance matrix for $\boldsymbol{\beta}$. Other choices are explored by Chipman et al. (2001) and Raftery et al. (1997), including a prior covariance matrix proportional to $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Several alternatives to the prior distribution (4) have been proposed in the literature. Two alternative formulations are:

- George and McCulloch (1993) chose a prior distribution for $\boldsymbol{\beta}$ that does not depend on $\sigma$.

$$p(\beta_j|\delta_j, \sigma) = p(\beta_j|\delta_j) \sim \begin{cases} N(0, \tau_j^2) & \text{if } \delta_j = 0, \\ N(0, c_j^2\tau_j^2) & \text{if } \delta_j = 1. \end{cases} \tag{5}$$

- Raftery et al. (1997) and Box and Meyer (1993) used a prior distribution similar to (4) but, when $\delta_j = 0$, the prior probability on $\beta_j$ is a point mass at 0. This is a limiting case of (4) with $c_j \to \infty$, $\tau_j \to 0$, and $c_j\tau_j$ fixed. It can be represented as

$$p(\beta_j|\delta_j, \sigma) \sim \begin{cases} \Delta(\beta_j) & \text{if } \delta_j = 0, \\ N(0, (c_j\tau_j\sigma)^2) & \text{if } \delta_j = 1, \end{cases} \tag{6}$$

  where the Dirac delta function $\Delta$ assigns probability 1 to the event $\beta_j = 0$. Formally, $\Delta(x)$ integrates to 1 and takes the value zero everywhere except at $x = 0$.

An important practical difference between priors (4)–(6) is the extent to which they allow analytic simplification of the posterior density as discussed in Section 2.3.

Prior distributions (4) and (6) are conjugate and allow analytic simplification of the posterior density. Prior (5) is nonconjugate, and requires additional computational techniques explored in Section 2.3.

### 2.2.2 Prior Distribution on the Subset Indicator Vector $\boldsymbol{\delta}$

A prior distribution must also be assigned to the subset indicator vector $\boldsymbol{\delta}$. Because $\boldsymbol{\delta}$ is a binary vector with $h$ elements, there are $2^h$ possible subsets of active effects. The prior distribution on $\boldsymbol{\delta}$ must assign probability to each subset. An initially appealing choice is to make each of the $2^h$ subsets equally likely. This is equivalent to prior independence of all elements of $\boldsymbol{\delta}$, and $p(\delta_j = 0) = p(\delta_j = 1) = 0.5$. In a screening context, this is implausible under the following widely accepted principles:

1. **Effect Sparsity:** Only a small fraction of all possible effects is likely to be active. Thus, the prior probability that $\delta_j = 1$ will be set to less than 0.5.
2. **Effect Hierarchy:** Lower-order effects are more likely to be active than higher-order effects. For example, linear main effects are more likely to be active than quadratic main effects or interaction effects.
3. **Effect Heredity:** Subsets should obey heredity of active effects. For example, a subset with an active $AB$ interaction but no $A$ or $B$ main effects may not be acceptable. Nelder (1998) referred to this as the "marginality principle".

The first two principles suggest that $P(\delta_j = 1)$ should be less than 0.5 for all $j = 1, \ldots, h$ and be smaller for higher-order effects. The third suggests that $p(\boldsymbol{\delta})$ should incorporate effect dependencies such as that proposed, for example, by Chipman (1996). There (and in this chapter) the probability that a given term is active or inactive is assumed to depend on its "parent" terms, typically taken to be those terms of the next lowest order from which the given term may be formed. A higher-order term such as $A_L B_Q$ could have parents $A_L B_L$ and $B_Q$ (the "typical" case), or have parents $A_L$ and $B_Q$.

To illustrate, consider a simple example with three factors, $A$, $B$, and $C$, and a model with three linear main effects and three linear $\times$ linear interactions. For clarity, elements of $\boldsymbol{\delta}$ are indexed as $\delta_A, \delta_{AB}$, and so on. Because all effects are linear or linear $\times$ linear interactions, the $L$ subscript is omitted from linear effects $A_L$, $B_L$, and $C_L$. The prior distribution on $\boldsymbol{\delta}$ is

$$\begin{aligned} p(\boldsymbol{\delta}) &= p(\delta_A, \delta_B, \delta_C, \delta_{AB}, \delta_{AC}, \delta_{BC}) \\ &= p(\delta_A, \delta_B, \delta_C) p(\delta_{AB}, \delta_{AC}, \delta_{BC} | \delta_A, \delta_B, \delta_C). \end{aligned} \tag{7}$$

The effect heredity principle motivates two simplifying assumptions. The first assumption is that terms of equal order are active independently of each other, given the activity of the lower-order terms. Then (7) becomes

$$p(\boldsymbol{\delta}) = p(\delta_A) p(\delta_B) p(\delta_C) p(\delta_{AB} | \delta_A, \delta_B, \delta_C) p(\delta_{AC} | \delta_A, \delta_B, \delta_C) p(\delta_{BC} | \delta_A, \delta_B, \delta_C).$$

Secondly, the activity of an interaction is assumed to depend only on the activity of those terms from which it is formed, so that we have

$$p(\boldsymbol{\delta}) = p(\delta_A)p(\delta_B)p(\delta_C)p(\delta_{AB}|\delta_A, \delta_B)p(\delta_{AC}|\delta_A, \delta_C)p(\delta_{BC}|\delta_B, \delta_C).$$

The prior distribution is specified by choosing marginal probabilities that each main effect is active, namely,

$$P(\delta_A = 1) = P(\delta_B = 1) = P(\delta_C = 1) = \pi \tag{8}$$

and by choosing the conditional probability that an interaction is active as follows:

$$P(\delta_{AB} = 1|\delta_A, \delta_B) = \begin{cases} \pi_0 \text{ if } (\delta_A, \delta_B) = (0, 0), \\ \pi_1 \text{ if one of } \delta_A, \delta_B = 1, \\ \pi_2 \text{ if } (\delta_A, \delta_B) = (1, 1). \end{cases} \tag{9}$$

Although, in principle, four probabilities could be specified in (9), the circumstances under which $(\delta_A, \delta_B) = (0, 1)$ and $(1, 0)$ can be distinguished are uncommon. In some applications, $\pi, \pi_0, \pi_1, \pi_2$ may vary for, say, effects associated with $A$ and with $B$. To keep notation simple, this straightforward generalization is not discussed further. Probabilities for the $AC$ and $BC$ interactions being active are defined in a similar way.

Effect sparsity and effect hierarchy are represented through the choice of hyperparameters $\pi, \pi_0, \pi_1, \pi_2$. Typically $\pi_0 < \pi_1 < \pi_2 < \pi < 0.5$. Section 2.3 provides details on the selection of these hyperparameters.

The choice of $\pi_0 = \pi_1 = 0, \pi_2 > 0$ in (9) allows an interaction to be active only if both corresponding main effects are active (referred to as *strong heredity*). The choice of $\pi_0 = 0, \pi_1 > 0, \pi_2 > 0$ allows an interaction to be active if one or more of its parents are active (*weak heredity*). Models obeying strong heredity are usually easier to interpret, whereas weak heredity may help the stochastic search to move around the model space by providing more paths between subsets. This is discussed further near the end of Section 3.3. Peixoto (1990) and Nelder (1998) argued in favor of strong heredity, because the models identified under this assumption are invariant to linear transformations of the factor levels. Chipman et al. (1997) suggested that, in the exploratory stages of data analysis, it may be desirable to relax the restrictions of strong heredity. They, instead, used the weak-heredity prior distribution with $\pi_1 < \pi_2$ to indicate a preference for strong-heredity models. A different choice of parameters in (9) was given by Box and Meyer (1993), in which $\pi_0 = 0, \pi_1 = 0, \pi_2 = 1$. An interaction would then be active only if all of its main effect parents are active, in which case it is forced to be active. This is referred to as an *effect forcing* prior distribution. Effect forcing greatly reduces the number of models under consideration, because activity of interactions is automatically determined by the activity of main effects.

Chipman (1996) and Chipman et al. (1997) adopted the convention that the parents of a term are those terms of the next lowest order. Thus $A_L B_Q$ has parents $A_L B_L$ and $B_Q$. Figure 2 illustrates this relationship. Each term that is a function of two factors has two parents. Probabilities such as (9) can be specified for higher-order terms. For those terms that involve only a single factor, one might

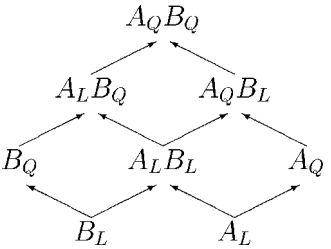



FIGURE 2. Ordering and inheritance relations among polynomial interactions.

specify

$$P(\delta_{AQ} = 1|\delta_{AL}) = \begin{cases} \pi_3 \text{ if } \delta_{AL} = 0, \\ \pi_4 \text{ if } \delta_{AL} = 1. \end{cases} \tag{10}$$

The hyperparameters $\pi_3, \pi_4$ would often be chosen as $\pi_3 = \pi_0, \pi_4 = \pi_2$.

If some factors are categorical with more than two levels, the corresponding effects are estimated via contrasts. If there are $l + 1$ levels, estimation of $l$ contrasts comparing each level with (say) level $l + 1$ are necessary. If $F$ is the categorical factor, and $\beta_{F1}, \ldots, \beta_{Fl}$ are the corresponding effects, then $\tau_{F1}, \ldots, \tau_{Fl}$ will indicate activity of these effects. Chipman (1996) suggested a prior distribution in which all of $\tau_{F1}, \ldots, \tau_{Fl}$ are either 0 or 1. This *effect grouping* allows either all or none of $\beta_{F1}, \ldots, \beta_{Fl}$ to be active.

## 2.3 The Posterior Distribution on the Subset Indicator $\boldsymbol{\delta}$

The joint posterior distribution of $\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}$ is proportional to the product of a likelihood from the linear model (1) and the prior distribution (2); that is,

$$p(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\beta}, \sigma, \boldsymbol{\delta})p(\boldsymbol{\beta}|\sigma, \boldsymbol{\delta})p(\sigma)p(\boldsymbol{\delta}). \tag{11}$$

Of primary interest in the subset selection problem is the marginal posterior distribution of the subset indicator $\boldsymbol{\delta}$:

$$\begin{aligned} p(\boldsymbol{\delta}|\boldsymbol{Y}) &\propto \iint p(\boldsymbol{Y}|\boldsymbol{\beta}, \sigma, \boldsymbol{\delta})p(\boldsymbol{\beta}|\sigma, \boldsymbol{\delta})d\boldsymbol{\beta}d\sigma \\ &= p(\boldsymbol{\delta}) \iint p(\boldsymbol{Y}|\boldsymbol{\beta}, \sigma, \boldsymbol{\delta})p(\boldsymbol{\beta}|\sigma, \boldsymbol{\delta})d\boldsymbol{\beta}p(\sigma)d\sigma \\ &= p(\boldsymbol{\delta})p(\boldsymbol{Y}|\boldsymbol{\delta}). \end{aligned} \tag{12}$$

The integral in (12) is evaluated either by MCMC methods (described in Section 3.1) or analytically. The result of integration, $p(\boldsymbol{Y}|\boldsymbol{\delta})$, is referred to as the *marginal likelihood* of $\boldsymbol{\delta}$, because it is the likelihood of $(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta})$ integrated over the prior distribution of $(\boldsymbol{\beta}, \sigma)$. Conjugate prior distributions (4) or (6) allow analytic integration. The nonconjugate prior distribution (5) of George and McCulloch (1993) requires MCMC integration.

For prior distributions (3) and (4), the marginal likelihood of $\boldsymbol{\delta}$ in (12) was given by George and McCulloch (1997) as

$$p(\boldsymbol{Y}|\boldsymbol{\delta}) \propto |\tilde{X}'\tilde{X}|^{-1/2}|\boldsymbol{D}_{\boldsymbol{\delta}}|^{-1}(\lambda\nu + \boldsymbol{S}_{\boldsymbol{\delta}}^2)^{(n+\nu)/2}, \tag{13}$$

where $\boldsymbol{D}_{\boldsymbol{\delta}}$ is a diagonal matrix with $j$th element equal to $\tau_j(1-\delta_j) + c_j\tau_j\delta_j$, and where

$$\boldsymbol{S}_{\boldsymbol{\delta}}^2 = \tilde{\boldsymbol{Y}}'\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}'\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{Y}},$$

$$\tilde{\boldsymbol{Y}} = \begin{bmatrix} \boldsymbol{Y} \\ 0 \end{bmatrix} \qquad \text{and} \qquad \tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{D}_{\boldsymbol{\delta}}^{-1} \end{bmatrix}.$$

The posterior density for the subset indicator $\boldsymbol{\delta}$ is then given by

$$p(\boldsymbol{\delta}|\boldsymbol{Y}) \propto p(\boldsymbol{\delta})p(\boldsymbol{Y}|\boldsymbol{\delta}) \equiv g(\boldsymbol{\delta}). \tag{14}$$

George and McCulloch (1997) also gave analytic results for a point mass mixture prior distribution (6).

In the above, the uninformative prior distribution on intercept $\beta_0$ is incorporated into computations by analysis of a centered response $Y_i - \bar{Y}$ and centered predictors. Thus, in computation, a model matrix $\boldsymbol{X}$ with no intercept column is actually used.

Whether the marginal posterior density on the subset indicator vector is calculated analytically (as in (13) and (14)) or via MCMC, there still remains the challenge of identifying subsets of active effects ($\boldsymbol{\delta}$ values) that have high posterior probability. The form of (13) suggests a similarity with least squares regression: the $\boldsymbol{S}_{\boldsymbol{\delta}}^2$ term acts as a residual sum of squares for a model relating $\tilde{\boldsymbol{Y}}$ to $\tilde{\boldsymbol{X}}$. Thus, in principle, efficient all-subsets search algorithms (Furnival and Wilson, 1974) could be used to identify subsets with high marginal likelihood (13). The marginal likelihoods would then be multiplied by the prior probability $p(\boldsymbol{\delta})$ to obtain the posterior density (14). The limitations of this approach have already been illustrated in Section 1: slow computation in large problems, and the vast majority of models with high $R^2$ will not obey strong or weak heredity. For these reasons, stochastic search methods based on MCMC techniques are a popular alternative. These are discussed in the next section.

## 3 Efficient Stochastic Search for Active Subsets

The focus of this section is on MCMC methods for sampling from $p(\boldsymbol{\delta}|\boldsymbol{Y})$, the posterior distribution of $\boldsymbol{\delta}$. In some cases, for example, in conjunction with the nonconjugate prior distribution (5), MCMC is also used to sample from $p(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}|\boldsymbol{Y})$.

It is impractical to evaluate the posterior probability (14) for all possible subsets of active effects due to the large number of possible models. Instead MCMC methods (either the Gibbs sampler or the Metropolis–Hastings algorithm) are used to draw samples from the posterior distribution. George and McCulloch (1997) discussed both the Metropolis–Hastings and Gibbs sampling algorithms for subset

selection and, in this context, MCMC algorithms may be thought of as stochastic search methods.

## 3.1 MCMC Sampling for the Subset Indicator Vector $\boldsymbol{\delta}$

The Gibbs sampler, used to sample from $p(\boldsymbol{\delta}|\boldsymbol{Y})$, starts with initial values of all parameters and then repeatedly draws values for each parameter conditional on all the others and the data. The steps below generate $K$ draws $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \ldots, \boldsymbol{\delta}^K$, that converge to the posterior distribution of $\boldsymbol{\delta}$:

1. Choose an initial value, $\boldsymbol{\delta}^0 = (\delta_1^0, \delta_2^0, \ldots, \delta_h^0)$,
2. For step $i = 1, \ldots, K$, do the following:
   - For $j = 1, \ldots, h$ in turn, draw a value $\delta_j^i$ from the distribution with density $p(\delta_j^i | \delta_1^i, \ldots, \delta_j^i, \delta_{j+1}^{i-1}, \ldots, \delta_h^{i-1}, \boldsymbol{Y})$.

Each draw is from a Bernoulli distribution. In drawing $\delta_j^i$, the $j$th component of $\boldsymbol{\delta}^i$, we condition on the most recently drawn values of all other components of $\boldsymbol{\delta}$. All values in the generated sequence $\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^K$ are treated as draws from the posterior distribution of $\boldsymbol{\delta}$.

For the nonconjugate prior distribution (5), George and McCulloch (1993) and Chipman et al. (1997) used the Gibbs sampler to simulate the joint posterior distribution of $(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta})$; that is, the above algorithm had, within step 2, an extra substep to draw values of $\beta_1, \ldots, \beta_h$ from $p(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma, \boldsymbol{Y})$ and a substep to draw values of $\sigma$ from $p(\sigma|\boldsymbol{\beta}, \boldsymbol{Y})$.

## 3.2 Estimation of Posterior Probability on $\boldsymbol{\delta}$ Using MCMC Output

In both the conjugate and nonconjugate cases just described, the most natural estimator of the posterior probability of a subset indicator vector $\boldsymbol{\delta}'$ is the observed relative frequency of $\boldsymbol{\delta}'$ among the $K$ sampled subsets $\mathcal{S} = \{\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \ldots, \boldsymbol{\delta}^K\}$; that is,

$$\hat{p}(\boldsymbol{\delta}'|\boldsymbol{Y}) = \frac{\sum_{i=1}^{K} I(\boldsymbol{\delta}^i = \boldsymbol{\delta}')}{K}. \tag{15}$$

Here, the indicator function $I(.)$ is 1 whenever its argument is true, and zero otherwise. A number of problems arise with the relative frequency estimate (15). First, it is prone to variability in the MCMC sample. Second, any model that is not in $\mathcal{S}$ has an estimated posterior probability of zero. Third, if the starting value $\boldsymbol{\delta}^0$ has very low posterior probability, it may take the Markov chain a large number of steps to move to $\boldsymbol{\delta}$ values that have high posterior probability. These initial *burn-in* values of $\boldsymbol{\delta}$ would have larger estimates of the posterior probability (15) than their actual posterior probability; that is, the estimate $\hat{p}(\boldsymbol{\delta}|\boldsymbol{Y})$ will be biased because of the burn-in. For example, with the simulated data described in Section 4.2, the first 100 draws of $\boldsymbol{\delta}$ have almost zero posterior probability. In a run of $K = 1000$

steps, the relative frequency estimate (15) of the posterior probability of these 100 draws would be $1/10$, whereas the actual posterior probability is nearly zero.

In conjugate settings such as (4) or (6), a better estimate of the posterior probability $p(\boldsymbol{\delta}|\boldsymbol{Y})$ is available. Instead of the relative frequency estimate (15), the analytic expression for the posterior probability (14) is used; that is,

$$p(\boldsymbol{\delta}'|\boldsymbol{Y}) = Cg(\boldsymbol{\delta}'), \tag{16}$$

provided that the normalizing constant $C$ can be estimated from the MCMC draws $\mathcal{S}$. Two methods for estimating $C$ are discussed below. The analytic estimate of posterior probability (16) is used throughout this chapter. Its use solves the problems of sampling variation and bias in estimating posterior probability due to burn-in described above.

The first approach to estimating the normalizing constant $C$ is to renormalize the probabilities for all unique subsets in the sampled set $\mathcal{S}$ so that they sum to 1. That is, all nonsampled subsets are assigned posterior probability 0. Let $\mathcal{U}$ be the set of unique $\boldsymbol{\delta}$ values in $\mathcal{S}$. The constant $C$ is then estimated by

$$\widehat{C} = \frac{1}{\sum_{i|\boldsymbol{\delta}^i \in \mathcal{U}} g(\boldsymbol{\delta}^i)}. \tag{17}$$

This estimate of $C$ and the resultant approximation to posterior probability (16) is commonly used when assigning posterior probability to models, and in summarizing features of the posterior distribution on $\boldsymbol{\delta}$. For example, the marginal probability of activity for a main effect $A$ can be calculated as the expected value of $\delta_A$, where $\delta_A$ is the component of $\boldsymbol{\delta}$ corresponding to main effect $A$. Based on the unique draws $\mathcal{U}$, the estimated posterior marginal probability would be

$$\Pr(\delta_A = 1|\boldsymbol{Y}) \approx \sum_{i \in \mathcal{U}} \delta_A^i \widehat{C} g(\boldsymbol{\delta}^i). \tag{18}$$

The use of (17) in marginal posterior probability (18) corresponds to the posterior probability that $A$ is active, conditional on the models visited by the MCMC run.

In some situations, such as when estimates of the posterior probability of a specific subset $\boldsymbol{\delta}$ or of groups of subsets are required, a second method of estimating $C$ can be used. Notice that the estimate $\widehat{C}$ in (17) will be biased upwards, because $\widehat{C} = C$ in (17) only if $\mathcal{U} = \mathcal{D}$, the set of all possible values of $\boldsymbol{\delta}$. If $\mathcal{U} \subset \mathcal{D}$, then $\widehat{C} < C$. A better estimate of $C$ can be obtained by a capture–recapture approach, as discussed by George and McCulloch (1997). Let the initial "capture set" $\mathcal{A}$ be a collection of $\boldsymbol{\delta}$ values identified before a run of the MCMC search; that is, each element in the set $\mathcal{A}$ is a particular subset. The "recapture" estimate of the probability of $\mathcal{A}$ is the relative frequency given by (15). The analytic expression (16) for the posterior probability of $\mathcal{A}$ is also available, and contains the unknown $C$. Let $g(\mathcal{A}) = \sum_{\boldsymbol{\delta} \in \mathcal{A}} g(\boldsymbol{\delta})$ so that $p(\mathcal{A}|\boldsymbol{Y}) = Cg(\mathcal{A})$. Then, by equating the two estimates, we have

$$Cg(\mathcal{A}) = \sum_{i=1}^{K} I(\boldsymbol{\delta}^i \in \mathcal{A})/K$$

and, on solving for $C$, a consistent estimator of $C$ is obtained as

$$\widehat{C} = \frac{1}{g(\mathcal{A})K} \sum_{k=1}^{K} I_{\mathcal{A}}(\boldsymbol{\delta}^k). \tag{19}$$

This technique is illustrated in Section 5.1.

Having seen that the observed frequency distribution of the sampler is useful in the estimation of the normalizing constant $C$, it is interesting to note that the frequencies are otherwise unused. Analytic expressions for the posterior probability are superior because they eliminate MCMC variability inherent in relative frequency estimates of posterior probabilities. In such a context, the main goal of the sampler is to visit as many different high probability models as possible. Visits to a model after the first add no value, because the model has already been identified for analytic evaluation of $p(\boldsymbol{\delta}|\mathbf{Y})$.

In later sections, (17) is used everywhere, except when estimating the posterior probability of all models visited by the MCMC sampler. In this case, (19) is used.

## *3.3 The Impact of Prior Distributions on Computation*

The choice of prior distributions discussed in Section 2.2 has an impact on both the ease of implementation of the MCMC algorithm, and its speed of execution. Specific issues include the number of linear algebra operations and the rate at which stochastic search methods can explore the space of all subsets.

The point mass prior on $\boldsymbol{\beta}$ in (6) reduces computation by dropping columns from the $\boldsymbol{X}$ matrix; that is, $\delta_j = 0$ implies $\beta_j = 0$, and the corresponding column is dropped. When a mixture of two normal distributions (4) or (5) is used instead, the $\boldsymbol{X}$ matrix is always of dimension $h + 1$.

Computations are also more efficient if the MCMC algorithm can sample directly from the marginal posterior distribution $p(\boldsymbol{\delta}|\boldsymbol{Y})$, rather than from the joint posterior distribution $p(\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma|\boldsymbol{Y})$. This efficiency occurs because fewer variables are being sampled. As mentioned at the end of Section 3.1, the marginal posterior distribution $p(\boldsymbol{\delta}|\boldsymbol{Y})$ is available in closed form when conjugate prior distributions on $\boldsymbol{\beta}$, as in (4) or (6), are used.

The prior distribution chosen for $\boldsymbol{\delta}$ affects the number of possible subsets of active effects that are to be searched. Figure 3 shows the relationship between the number of possible subsets and the number of factors when linear main effects and linear $\times$ linear interactions are considered under four different priors. For $f$ factors, the number of effects under consideration is

$$h = f + \binom{f}{2} = (3f + f^2)/2.$$

The number of possible subsets of active effects will be $2^h$ if all possible subsets of active effects are considered. Hence a log base 2 transformation is used on the vertical axis of Figure 3. Strong and weak heredity reduce the number of subsets

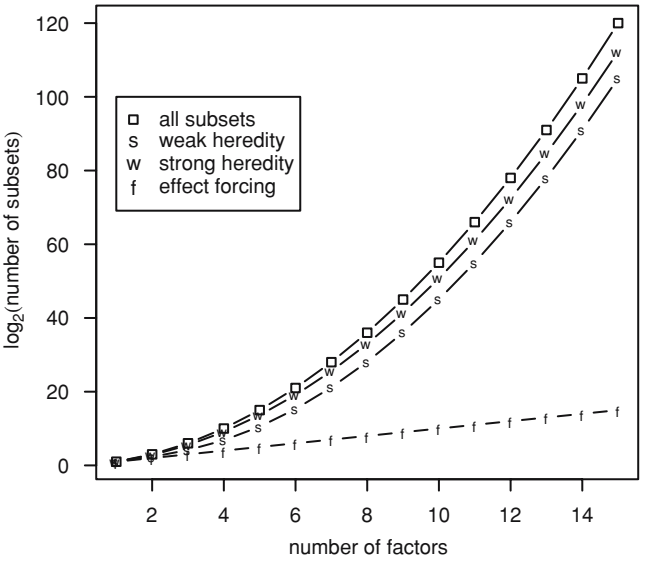



FIGURE 3. Relationship of the number of possible subsets of active effects ($\log_2$ scale) and the number of factors for models with linear $\times$ linear interactions and linear main effects only.

that can occur. The effect forcing of Box and Meyer (1993) yields far fewer subsets, because activity of interactions is automatically determined by activity of main effects. With effect forcing, the posterior probability can easily be calculated for every possible subset, rather than via stochastic search methods.

Prior distributions on $\boldsymbol{\delta}$ that enforce heredity can also affect the manner in which stochastic search algorithms move around the subset space. The Gibbs sampling algorithm described in Section 3.1 updates one element of $\boldsymbol{\delta}$ at a time, conditional on the values of all other elements; that is, the algorithm randomly adds or drops a single effect from the subset at each step. This is a common approach used by many stochastic search algorithms. Strong heredity restricts the number of paths between subsets. For example, to move from the subset $\{A, B, C, AB, AC\}$ to the subset $\{A\}$, it would be necessary to drop $AB$ before dropping $B$ and to drop $AC$ before dropping $C$. Weak heredity, which allows subsets like $\{A, C, AB, AC\}$, provides more paths between subsets, enabling the stochastic search to move more freely around the subset space.

Prior distributions that enforce effect grouping, such as those for indicator variables, also have an impact on computation. Unlike the Gibbs sampler, where changing a single element $\delta_j$ affects only one element $\beta_j$ of $\boldsymbol{\beta}$, with grouped effects, a change in one element of $\boldsymbol{\delta}$ implies a change to several elements of $\boldsymbol{\beta}$. Because updating formulae are used in calculating how changes to $\boldsymbol{\beta}$ affect the posterior distribution of $\delta$, a simultaneous change to several elements of $\boldsymbol{\beta}$ may entail more complicated updates.

# 4 Selection of Hyperparameters of the Prior Distribution

Choices of values for the hyperparameters, $\nu, \lambda, c_j, \tau_j$, of the prior distributions (3) and (4) are now discussed, with an emphasis on automatic methods that use simple summaries of the observed data. Many of these suggestions have been made by Chipman et al. (1997) and Chipman (1998). Some suggestions relating to the choice of $c$ and $\tau$ are new.

## *4.1 Hyperparameters for the Prior Distribution on $\sigma$*

I propose that the hyperparameters for the prior distribution (3) on $\sigma^2$ are chosen so that the mean and 99th quantile of the distribution are consistent with the observed values of the response. The prior expected value of $\sigma^2$ is

$$\mathrm{E}(\sigma^2) = \frac{\lambda\nu}{\nu - 2}, \quad \text{for } \nu > 2,$$

which suggests that $\lambda$ should be chosen near the expected residual variance. In the absence of expert knowledge, some fraction of the sample variance of the response, $s^2 = \sum_{i-1}^{n}(y_i - \bar{y})^2/(n-1)$ could be used to choose $\lambda$. Chipman et al. (1997) proposed

$$\lambda = s^2/25 = \left[\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)\right]/25.$$

This represents the prior belief that the residual standard deviation will be roughly $1/5$ of the standard deviation of the response.

As was observed after (3) in Section 2.2, the hyperparameter $\nu$ can be thought of as the amount of information about $\sigma$ arising from a sample with $\nu$ observations and sample variance $\lambda$. The parameter $\nu$ controls the length of the right tail of the prior distribution (3) for $\sigma$. Larger values of $\nu$ imply a prior distribution that is more tightly centered around $\lambda$. Various quantiles for an inverse gamma distribution with $\lambda = 1$ are given in Table 5 and it can be seen that the distribution has quite a long right tail for $\nu \leq 5$. A sufficiently diffuse prior distribution may be selected by choosing $\nu$ so that the upper tail (say the 99th percentile) is not far below $s^2$. The choice of $\nu = 5$ would place the 99th percentile of the prior distribution of $\sigma^2$ at $9.02\lambda$, for example. Combining this with the proposed $\lambda = s^2/25$ gives the 99th prior quantile of the distribution of $\sigma$ as $9.02s^2/25 = 0.36s^2$. Smaller values for $\nu$ are possible (for example, the value $\nu = 1.5$ was used by Chipman et al., 1997), although they can lead to unreasonably long tails, because the variance of an Inverse Gamma distribution is not defined for $\nu \leq 4$. In general, one would choose

$$\nu = 5 \quad \text{or select from Table 5.}$$

To sum up the section, the choices $\lambda = s^2/25$ and $\nu = 5$ are recommended.

TABLE 5. The mean and various quantiles of Inverse Gamma distributions with $\lambda = 1$ and $\nu = 1, \ldots, 10$. For other values of $\lambda$, multiply the table entries by $\lambda$ to obtain the appropriate quantile

| $\nu$ | Mean | 0.01 | 0.1 | 0.5 | 0.9 | 0.99 |
|---|---|---|---|---|---|---|
| 1 | — | 0.15 | 0.37 | 2.2 | 63.33 | 6365 |
| 2 | — | 0.22 | 0.43 | 1.44 | 9.49 | 99.50 |
| 3 | 3 | 0.26 | 0.48 | 1.27 | 5.13 | 26.13 |
| 4 | 2 | 0.30 | 0.51 | 1.19 | 3.76 | 13.46 |
| 5 | 1.67 | 0.33 | 0.54 | 1.15 | 3.10 | 9.02 |
| 6 | 1.5 | 0.36 | 0.56 | 1.12 | 2.72 | 6.88 |
| 7 | 1.4 | 0.38 | 0.58 | 1.10 | 2.47 | 5.65 |
| 8 | 1.33 | 0.40 | 0.60 | 1.09 | 2.29 | 4.86 |
| 9 | 1.29 | 0.42 | 0.61 | 1.08 | 2.16 | 4.31 |
| 10 | 1.25 | 0.43 | 0.63 | 1.07 | 2.06 | 3.91 |

## 4.2 *Hyperparameters for the Prior Distribution on $\beta$*

Prior distribution (4) for $\boldsymbol{\beta}$ is defined by the hyperparameters $c_j$ and $\tau_j$ ($j = 1, \ldots, h$). In choosing these hyperparameters, it is helpful to recall from (4) that the coefficient $\beta_j$ associated with an inactive contrast has standard deviation $\sigma\tau_j$ and, if the contrast is instead active, then $\beta_j$ has a standard deviation that is $c_j$ times larger. As mentioned earlier in Section 2.2, intercept $\beta_0$ is always present in the model, so taking $c_0 = 1$, $\tau_0 \to \infty$ gives a flat prior density for $\beta_0$.

Box and Meyer (1986) suggested the choice $c_j = 10$, for $j = 1, \ldots, h$, thus separating large and small coefficients by an order of magnitude. George and McCulloch (1997) suggested the following technique for the choice of $\tau_j$. If the $j$th contrast is inactive, even large changes in the contrast coefficients in the $j$th column of $\boldsymbol{X}$ should produce only a small change in the mean response. Such a small change in the mean response (say $\Delta Y$) could be considered to be similar in size to the residual standard deviation $\sigma$. A small coefficient $\beta_j$ has standard deviation $\sigma\tau_j$ and its value will lie in the range $0 \pm 3\sigma\tau_j$ with very high probability. Denote by $\Delta X_j$ the difference between the largest and smallest element in the $j$th column of $\boldsymbol{X}$. Then, when the contrast coefficient changes by $\Delta X_j$, the mean response is unlikely to change by more than $3\sigma\tau_j\Delta X_j$. On solving $\Delta Y \approx \sigma = 3\sigma\tau_j\Delta X_j$, we obtain $\tau_j = 1/(3\Delta X_j) = 1/3(\max(X_j) - \min(X_j))$. In two-level designs, in which contrast coefficients are coded as $+1$ and $-1$, $\Delta X_j = 2$. In summary, the default choice of the hyperparameter values is, for $j = 1, \ldots, h$,

$$c_j = c = 10, \quad \tau_j = \frac{1}{3(\max(X_j) - \min(X_j))}. \tag{20}$$

An alternative choice, labelled (21), is discussed later in this section.

The use of minimum and maximum coefficients of each contrast makes the method invariant to rescalings of the contrasts. In the glucose example, contrasts are coded with quite different ranges. This will not have an impact on the analysis because the definition of large and small effects is adjusted accordingly.

The subset selection procedure can be sensitive to the choice of $\tau_j$ (see Chipman et al., 1997 and George and McCulloch, 1993). Equation (20) captures the relative magnitudes of the $\tau_j$ for different variables, but the overall magnitude may need tuning. Box and Meyer (1993) and Chipman et al. (1997) proposed methods for tuning based on runs of the search algorithm. A faster alternative (Chipman, 1998) based on predictive distributions is discussed here. For any given subset $\boldsymbol{\delta}$ and associated model, the posterior mean of $\boldsymbol{Y}$ for a given $\boldsymbol{X}$ may be calculated using the posterior mean of $\boldsymbol{\beta}$. The magnitude of $\tau_j$ determines the degree of shrinkage (toward zero) in the estimate of the coefficient $\beta_j$ in a manner similar to ridge regression. A simple way to assess the impact of the $\tau_j$ is to see how the predictions (values of the posterior mean of $\boldsymbol{Y}$) change for various multiples, $r$, of default choices $\tau_1, \ldots, \tau_h$, where $r$ ranges over the interval (1/10,10), for a single given model. A good $\tau_j$ value would be the smallest value that does not shrink the posterior predictions too far from the least squares predictions.

The posterior mean for $\boldsymbol{\beta}$ conditional on a particular subset of active effects $\boldsymbol{\delta}$ is obtained by integration of $p(\boldsymbol{\beta}, \sigma|\boldsymbol{Y}, \boldsymbol{\delta})$ with respect to $\sigma$ to obtain $p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\delta})$, and by calculation of an expectation of $\boldsymbol{\beta}$ with respect to this distribution. The resultant posterior mean is shown by George and McCulloch (1997) to be

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\delta}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_{\boldsymbol{\delta}}^{-2})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

where $\boldsymbol{D}_{\boldsymbol{\delta}}$ is a diagonal matrix with elements $\tau_j(1 - \delta_j) + \tau_j c_j \delta_j$.

To illustrate the impact of the choice of the $\tau_j$, two examples are considered; first, a simulated example and then the glucose data.

Hamada and Wu (1992) and Chipman et al. (1997) discussed a simulated screening experiment using the 12-run Plackett–Burman design shown in Table 6. The 11 factors are labeled $A$–$K$ and each has two levels coded as $+1$ and $-1$. The response was simulated from the true model

$$Y = A + 2AB + 2AC + \varepsilon, \quad \varepsilon \sim N(0, \sigma = 0.25).$$

In this model, factor $A$ has an active main effect and there are active interactions between $A$ and $B$ and between $A$ and $C$, and the remaining factors $D$–$K$ are inactive.

Figure 4 shows the predictions (that is, the posterior mean of $\boldsymbol{Y}$) in this simulated example. There are 12 design points (runs) and, for a model involving two-factor interactions, there are $h = 11 + 55 = 66$ potentially active effects. The default choices of $\tau_j$, $j = 1, \ldots, 66$, given by (20), are all equal to $\tau = 1/6$ and are multiplied by values of $r$ from (1/10, 10). Stepwise regression identified correctly the subset of active effects, $A$, $AB$, $AC$. Main effects for $B$ and $C$ were subsequently included so that the subset ($A$, $B$, $C$, $AB$, $AC$) obeys strong heredity. This set of effects was then used in the calculations of the posterior mean (predicted response). The value 1.0 on the horizontal axis in Figure 4 denotes the default choice for $\tau$. Both the observed $Y_i$ values ($\bullet$) and predictions based on the least squares estimate $\widehat{\boldsymbol{\beta}}$ ($\circ$) are shown on the right side of the plot. In this example, the default choice for $\tau$ seems quite reasonable, as any smaller

TABLE 6. Plackett–Burman 12-run design with simulated response data

| Factor | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | B | C | D | E | F | G | H | I | J | K | Response |
| + | + | − | + | + | + | − | − | − | + | − | 1.058 |
| + | − | + | + | + | − | − | − | + | − | + | 1.004 |
| − | + | + | + | − | − | − | + | − | + | + | −5.200 |
| + | + | + | − | − | − | + | − | + | + | − | 5.320 |
| + | + | − | − | − | + | − | + | + | − | + | 1.022 |
| + | − | − | − | + | − | + | + | − | + | + | −2.471 |
| − | − | − | + | − | + | + | − | + | + | + | 2.809 |
| − | − | + | − | + | + | − | + | + | + | − | −1.272 |
| − | + | − | + | + | − | + | + | + | − | − | −0.955 |
| + | − | + | + | − | + | + | + | − | − | − | 0.644 |
| − | + | + | − | + | + | + | − | − | − | + | −5.025 |
| − | − | − | − | − | − | − | − | − | − | − | 3.060 |

multiples would shrink the posterior mean away from the data (•) and the least squares estimates (○).

In Figure 5(a), a similar plot is given for the glucose data, where the model contains the subset of active effects $B_L$, $H_L$, $B_Q$, $H_Q$, $B_L H_L$, $B_L H_Q$, $B_Q H_L$, $B_Q H_Q$. Close inspection of this plot reveals a problem: the posterior mean of $Y_i$ converges

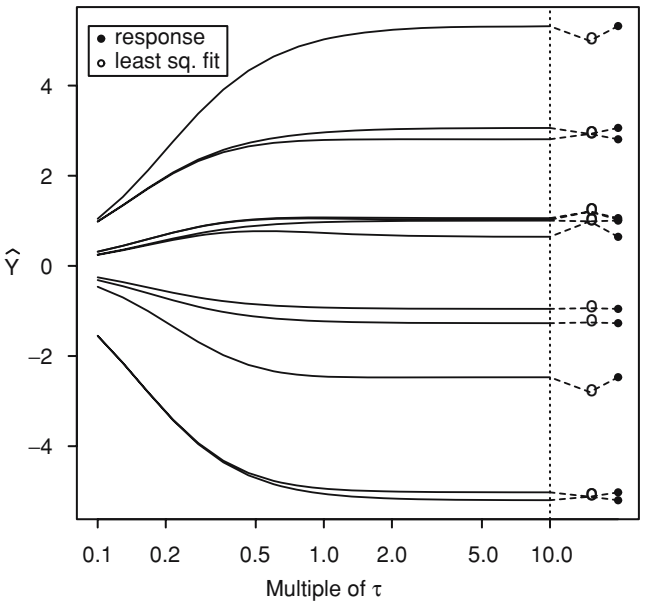



FIGURE 4. Predicted response at each of the 12 design points (two coinciding) in Table 6 for $c = 10$ and various multiples of the default value for $\tau = 1/6$.

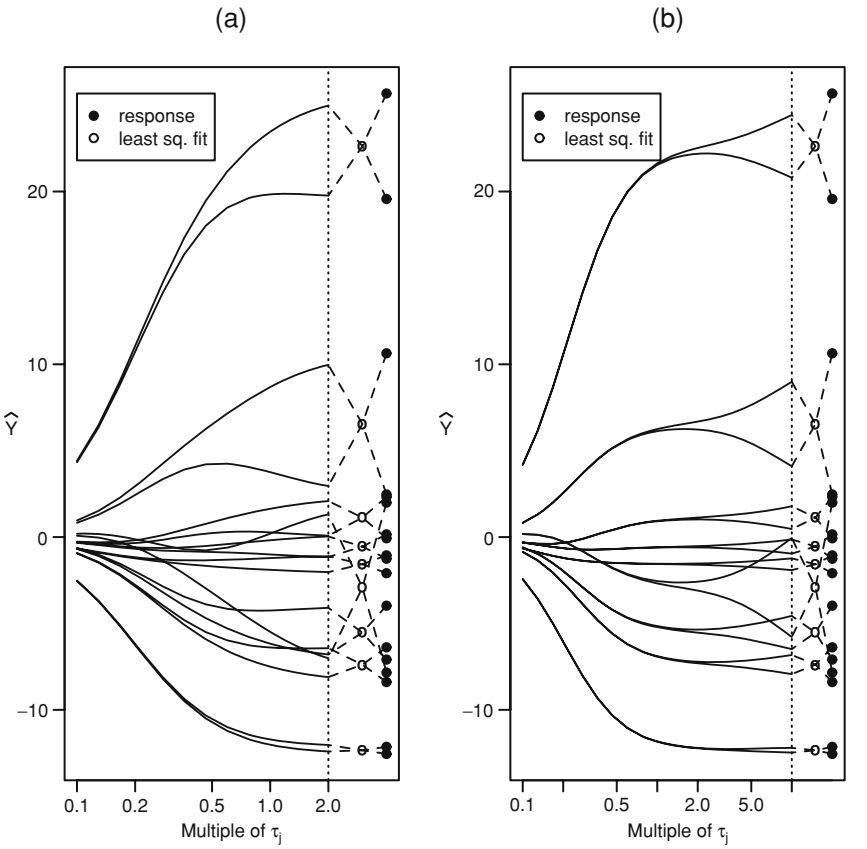



FIGURE 5. Glucose data: Predicted response at each of the 18 design points of Table 2 for various multiples of the default $\tau_j$; default values of $c_j$ and $\tau_j$ are calculated (a) from equations (20) and (b) from equations (21). The model used has active effects $B_L$, $H_L$, $B_Q$, $H_Q$, $B_L H_L$, $B_L H_Q$, $B_Q H_L$, $B_Q H_Q$.

to the corresponding data value (•) rather than to the least squares estimate (○). This is somewhat unexpected: one might suppose that as $\tau_j$ is increased, so that less prior information is used, the posterior means would approach the least squares estimates. The posterior means converge instead to the data because, in addition to the intercept and eight active effects ($B_L, \ldots, B_Q H_Q$), there are $113 - 8 = 105$ other "inactive" effects included in the model. The estimates of these are heavily shrunk toward zero, but not set exactly to 0, because the prior distribution (4) specifies that inactive effect $\beta_j$ has prior standard deviation $\tau_j \sigma$. Some of the residual variation not captured by the eight active effects is absorbed by the 105 inactive effects rather than being captured in the error term $\varepsilon$.

The tendency of inactive effect estimates to capture residual variation suggests an alternative strategy for choosing values for $c_j$ and $\tau_j$ in experiments with very

large numbers of candidate effects, namely, to reduce the prior magnitude of an inactive effect. This can be accomplished by multiplying $c_j$ in (20) by, say 10, and by dividing $\tau_j$ in (20) by 10, giving alternative default choices of $c_j$ and $\tau_j$ of

$$c_j = 100, \qquad \tau_j = \frac{1}{30 \times \text{range}(\boldsymbol{X}_j)}. \tag{21}$$

Because an active effect has prior standard deviation $\sigma c_j \tau_j$, this modification has no effect on the distribution of the active effects. The prior standard deviation $\sigma \tau_j$ of a small effect has been shrunk by a factor of 10, approaching the point mass prior distribution of Raftery et al. (1997) and Box and Meyer (1993).

The use of multiples of the default $\tau_j$ in (21) for the glucose example produces Figure 5(b). With a multiplier of 1–2, the posterior means first approach the least squares estimates. For multiples in this range, the inactive effects are still shrunk quite close to zero and unable to absorb residual errors. Only when the multiplier of $\tau$ is quite large (such as a factor of 10 times the default) do the fitted values converge to the data points.

To summarize, I recommend that (20) be used to choose $c_j$, $\tau_j$, unless a very large number of inactive effects are anticipated. In that case, (21) should be used. In either case, a check of the effect of scaling the $\tau_j$ values via a plot such as Figure 4 is helpful.

## *4.3 Hyperparameters for the Prior Distribution on Subset Indicator $\boldsymbol{\delta}$*

Box and Meyer (1986) examined several published analyses of experiments and concluded that, in these, between 10% and 30% of main effects were identified as active, with an average of 20% active effects. Similar arguments may be made for screening experiments, with some modification for interactions and higher-order effects. The suggestions made here utilize calculations for the expected number of active effects. Such expectations should be easier to specify than prior probabilities.

First, a simplified method of choosing the hyperparameters is described following Bingham and Chipman (2002). The probability of an active interaction is assumed to be proportional to $\pi$, the probability of an active linear main effect, with the size of the proportionality constant dependent on which parents are active. Thus (9) becomes

$$P(\delta_{AB} = 1 | \delta_A, \delta_B) = \begin{cases} \pi a_0 & \text{if } (\delta_A, \delta_B) = (0, 0), \\ \pi a_1 & \text{if one of } \delta_A, \delta_B = 1, \\ \pi a_2 & \text{if } (\delta_A, \delta_B) = (1, 1), \end{cases} \tag{22}$$

with proportionality constants $a_0, a_1, a_2$ and (10) becomes

$$P(\delta_{AQ} = 1 | \delta_{AL}) = \begin{cases} \pi a_3 & \text{if } \delta_{AL} = 0, \\ \pi a_4 & \text{if } \delta_{AL} = 1, \end{cases} \tag{23}$$

with proportionality constants $a_3$ and $a_4$. Bingham and Chipman (2002) chose $(a_0, a_1, a_2) = (0.01, 0.5, 1.0)$. For quadratic effects one might choose $(a_3, a_4) = (0.01, 1.0)$. Such choices reduce the selection of a prior distribution on $\boldsymbol{\delta}$ to the specification of the value of a single hyperparameter $\pi$. The rationale for these choices is as follows: $a_2 = a_4 = 1.0$ corresponds to the belief that, if all parents of an effect are active, then that effect has the same probability of activity as each of its parents. At the other extreme, with $a_0 = a_3 = 0.01$, when none of the parents of an effect are active, then it is highly unlikely that the effect will be active. The remaining choice, $a_1 = 0.5$, corresponds to weak heredity in that an effect with one out of two active parents has some chance of activity, but it should be smaller than if both parents were active.

The value of the hyperparameter $\pi$ may be chosen by considering the prior expected number of active effects. Illustrative calculations are now given for a full second-order model with $f$ factors, and for subsets of active effects that include linear and quadratic main effects and linear $\times$ linear interactions. Thus, the full model contains $f$ linear effects, $f$ quadratic effects, and $\binom{f}{2}$ linear $\times$ linear interaction effects. Prior probabilities on the subsets being active have the form of (22) and (23) above. A straightforward extension of the calculations of Bingham and Chipman (2002) yields an expected number of active effects as

$$E(\#\text{ active effects}) = f\pi + f\pi\{(1-\pi)a_3 + \pi a_4\} + \pi \binom{f}{2} \{a_0 + 2\pi(a_1 - a_0) + \pi^2(a_0 - 2a_1 + a_2)\}. \quad (24)$$

The two terms on the first line of (24) represent the expected number of active linear and quadratic main effects; the second line gives the expected number of active linear $\times$ linear interaction effects.

This simple expression is invaluable for elicitation of a prior distribution: instead of specifying the probability of activity for a variety of different effects, an expected number of effects can be specified, along with values of $a_0, \ldots, a_4$ (for example, the defaults suggested below (23)). The expression for the expected number of active effects (24) is then solved for $\pi$. A data analyst could also experiment with alternative values of $a_0, \ldots, a_4$ and see the impact of these choices in terms of the expected number of linear effects and interactions.

Choices of $\pi$ yielding 2, 4, and 6 expected active effects for the simulated example of Section 4.2 are shown in Table 7. In this example, there were 11 linear effects and 55 linear $\times$ linear interaction effects. The columns of the table show

TABLE 7. Expected numbers of active effects, corresponding hyperparameter $\pi$, and breakdown by linear main effects and linear $\times$ linear interaction effects, for the simulated experiment with 11 factors

| E(# active effects) | $\pi$ | E(# linear effects) | E(# linear $\times$ linear int.) |
|---|---|---|---|
| 2 | 0.113 | 1.245 | 0.754 |
| 4 | 0.185 | 2.040 | 1.960 |
| 6 | 0.243 | 2.674 | 3.326 |

the different number of effects expected to be active for different values of $\pi$ and the default values for $a_0, \ldots, a_4$ given below (23).

# 5 Examples

Stochastic simulation methods for fitting Bayesian models are now discussed and illustrated using the two examples that were described earlier in the chapter.

## *5.1 A Simulated Experiment*

The simulated screening experiment was introduced in Section 4.2 and the data are shown in Table 6. All 11 factors ($A$–$K$) are set at two levels, and so only linear main effects and linear $\times$ linear interaction effects are considered. All corresponding contrast coefficients are $+1$ or $-1$. The weak heredity prior distribution on $\boldsymbol{\delta}$, (22), is calibrated with $\pi = 0.185$ so that there are four expected active effects. The default choice (20) of $\tau_j$ and $c_j$ yields $\tau_j = \tau = 1/3(1-(-1)) = 1/6$ and $c_j = c = 10$, for all $j$. For the prior distribution on $\sigma$, default choices are $\lambda = s^2/25 = 10.01/25 = 0.40$ and $\nu = 5$.

One thousand draws from the posterior distribution (14) were collected via the Gibbs sampler. The probability that each effect is active is plotted in Figure 6(a), as a vertical line. This plot is quite similar to that of Figure 1 in Section 1.1: details are given below. It is quite clear that the active effects ($A$, $AB$, $AC$) are well identified by the algorithm and that all other effects are correctly identified as inactive. The marginal posterior probabilities plotted in Figure 6 have been calculated analytically using (17) and (18) rather than the relative frequency estimates based on (15).

The joint posterior probability distribution on $\boldsymbol{\delta}$ is also informative. The true model ($A$, $AB$, $AC$) dominates, with 50.1% of posterior probability. The two next most probable models each have posterior probability of about 3%. Each involves the addition of either the $B$ or $C$ linear effects to the most probable model. Posterior probabilities reported here are normalized (using (17)) to sum to 1 over all distinct models visited by the 1000 Gibbs sampler draws.

An important feature of any subset selection procedure is that it should be able to identify the situation where no effects are active. In order to explore this, the 12 values in the response vector $\boldsymbol{Y}$ were randomly shuffled in the order 6, 2, 7, 8, 9, 1, 4, 12, 5, 3, 11, 10, and rows of $\boldsymbol{X}$ were not changed. The analysis was re-run and the marginal probabilities plotted in Figure 6(b). Although a few factors have some probability of activity, there is nothing quite as convincing as in the original analysis, see below.

To explore the sensitivity of the algorithm to a variety of hyperparameter choices, various combinations of $\pi$, $\tau$, and $c$ were considered. The three values of $\pi$ given in Table 7 were used, giving 2, 4, and 6 expected active effects. Six combinations of $\tau$ and $c$ were explored. The first three choices were $c = 10$ and $\tau = (0.5, 1, 1.5)/6$. These are close to the default hyperparameter choice (20) of $c = 10$, $\tau = 1/6$. The second three choices were $c = 100$ and $\tau = (0.05, 0.1, 0.15)/6$. These are close

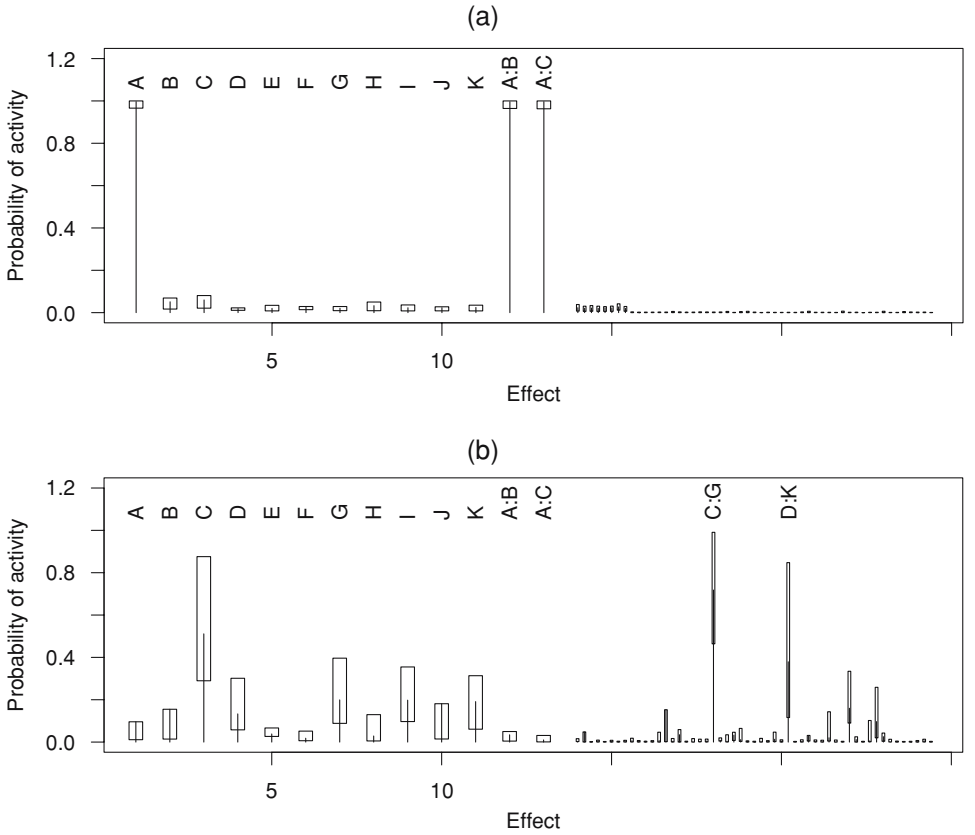



FIGURE 6. (a) Marginal probability of activity for each effect for the simulated screening experiment. The lines correspond to hyperparameter settings that give four expected active effects ($\pi = 0.185, a_0 = 0.01, a_1 = 0.5, a_2 = 1.0$), default choice of $\tau$ and $c = 10$. Rectangles represent extremes over 2, 4, 6 prior expected effects and six $(c, \tau)$ multipliers of (1, .5), (1, 1), (1, 1.5), (10, .05), (10, .1), (10, .15). (b) Same plot, except that the values in the response vector have been permuted.

to the default hyperparameter choice (21) of $c = 100$, $\tau = 0.1/6$. The rectangles in Figure 6 represent the range of posterior probabilities over the 18 combinations of hyperparameters. Figure 6(a) shows that there is minimal sensitivity to the hyperparameters, as the boxes are narrow and most probabilities are near 0 or 1. In Figure 6(b), with no signal, there is considerably more uncertainty and no effects are identified "active" as clearly as in panel (a). Three effects have posterior probabilities of activity of over 50% (C, CG, DK) under at least one of the 18 prior settings. The CG effect in particular appears quite active, with a posterior probability nearly as high as that of $B^2H^2$ in Figure 1 for the glucose experiment. Unlike the glucose experiment, there is no consistent set of variables appearing in many active effects.

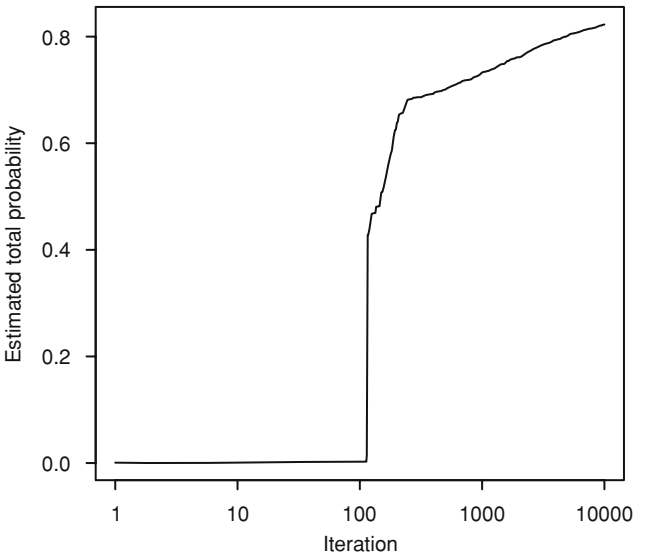



FIGURE 7. Estimated cumulative posterior probability of the distinct models visited, up to each iteration, for the simulated screening experiment.

One issue with stochastic simulation methods is for how long a simulation should be run. This can be addressed, in part, by estimating the posterior probability of all models (subsets) visited so far. To estimate this probability in the above example, the normalizing constant $C$ was estimated via the capture–recapture method (19) discussed in Section 3.2. The "capture set" $\mathcal{A}$ was chosen as the first 1000 draws from a run of 10,000 iterations. The capture set $\mathcal{A}$ contained 364 different values of $\boldsymbol{\delta}$ (that is, 364 distinct subsets of active effects). The other 636 $\boldsymbol{\delta}$ values visited in the first 1000 iterations were duplicates of these 364 and were not included in $\mathcal{A}$. In the remaining 9000 iterations, 73% of the $\boldsymbol{\delta}$ values visited were contained in $\mathcal{A}$. Thus in (19), $\sum_{i=1}^{K} I_{\mathcal{A}}(\boldsymbol{\delta}^k)/K = 0.73$. After calculation of the estimated normalizing constant via (19), it is estimated that, by the end of 10,000 iterations, the models visited account for 82% of the posterior probability. The estimated cumulative posterior probability of models visited is graphed in Figure 7. The algorithm takes approximately 100 iterations to identify a high-probability indicator vector $\boldsymbol{\delta}$. By 1000 iterations there appear to be very few high probability models that have not been visited, because the slope of the curve has decreased (and continues to decrease, because the horizontal axis is on a $\log_{10}$ scale). There is little advantage in running many more than 1000 iterations for this problem.

Figure 7 also illustrates the "burn-in" problem with relative frequency estimates of posterior probability which was discussed in Section 3.2. The first 100 iterations of the algorithm visit improbable subsets, so a relative frequency estimate of posterior probability (15) will place too much probability on these 100 subsets.

## 5.2 Glucose Experiment

The main results of the analysis of the glucose data have already been presented in Section 1.1. Details of the hyperparameter choices, robustness calculations, and estimation of the total probability visited by the MCMC sampler are given here.

The hyperparameters in the prior distributions of $\sigma^2$ and $\boldsymbol{\beta}$ are set as follows for this example, unless otherwise indicated, $\nu = 5$, $\lambda = s^2/25 = 101.2282/25 = 4.049$; $c_j = c = 100$ and $\tau_j$ are specified according to (21). The exact values of $\tau_j$ vary with effect index $j$, because the factor levels have different ranges. As discussed in Section 4.2, $c = 10$ seems to allow too much flexibility for the inactive effects to capture residual error. Calibration of $\pi$ via an expected number of effects is difficult, because effects of so many types (linear, quadratic, linear $\times$ linear, linear $\times$ quadratic, quadratic $\times$ quadratic) are present but calibration can be carried out using only the expected number of linear and quadratic main effects and linear $\times$ linear interaction effects. There are 8 possible linear, 7 quadratic, and 28 linear $\times$ linear interaction effects, for a total of 43 possible effects. The choice of $\pi = 0.2786$ gives 5 effects expected to be active out of the 43. The inclusion of higher-order interactions will raise this expectation, but not by much, because all their parents are of at least second order and are unlikely to be active.

A single run of the MCMC sampler was used, with 2500 iterations. The posterior probabilities of the models listed in Table 4 are normalized so that all subsets visited have total probability 1.0 of being active; that is, estimate $\widehat{C}$ from (17) is used in conjunction with analytic expression (16) for the posterior probability on $\boldsymbol{\delta}$.

In a study of robustness, choices of $\pi = 0.1486$, 0.2786, and 0.3756 were considered, giving 2, 5, and 8 expected effects (considering up to linear $\times$ linear effects only, as above). Six combinations of $c$ and $\tau_j$ settings were used, as in the last example, with three $\tau_j$ values set at 0.5, 1.0, and 1.5 times the defaults in (20) with $c = 10$ and (21) with $c = 100$. The vertical lines in Figure 1 correspond to the default choices given at the start of this section, and the rectangles correspond to ranges over the 18 different hyperparameter settings.

The overwhelming conclusion from both Figure 1 and Table 4 is that high-order interactions between $B$ and $H$ are present. As Chipman et al. (1997) mentioned, this could well be due to the choice of the original factors: products of volume ($B$) and dilution ($H$) might give some absolute amount of blood in the sample. A transformation might eliminate the need for higher-order effects.

A long run of 50,000 iterations was carried out to estimate the posterior probability that subsets visited by the MCMC algorithm are active, via the capture–recapture method (19). The first 500 iterations determined a capture set $\mathcal{A}$, and the remaining 49,500 iterations were used to estimate the total posterior probability of subsets visited. Figure 8 shows that just slightly over 40% of the probability is visited by 50,000 iterations. The steadily increasing total probability in Figure 8 indicates that posterior probability is spread over a very large number of models. Such a diffuse posterior distribution requires many iterations of the Gibbs sampler to capture a significant fraction of the total probability. The small portion of posterior probability visited by the search during the first 2500 iterations (roughly

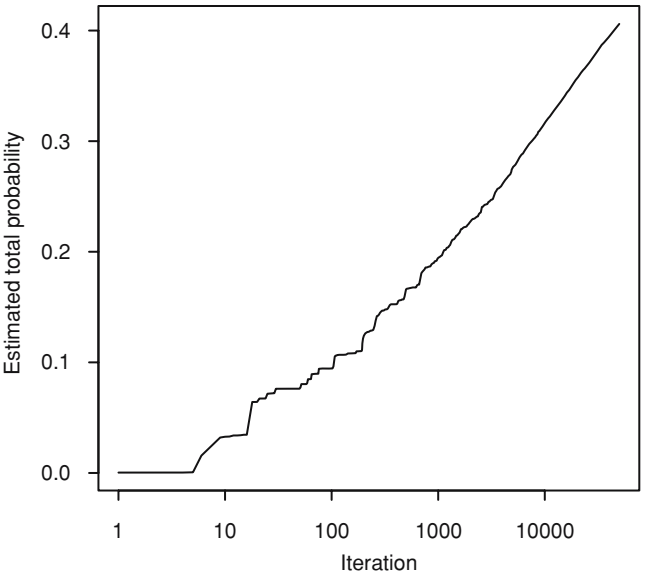



FIGURE 8. Estimated cumulative probability of the distinct models visited for the glucose experiment, for up to 50,000 iterations.

25%) implies that true subset probabilities are likely to be $1/4$ the values given in Table 4 (posterior probabilities in the table are normalized to sum to 1). What is perhaps more important is the relative size of the probabilities, given that the posterior distribution is quite diffuse. In this problem, more information may be obtained from marginal posterior distributions on individual $\delta_j$ than from the joint posterior distribution on $\boldsymbol{\delta}$.

## 6 Prior Distributions for Design

Bayesian methods have often proved useful for design of experiments, especially in situations in which the optimal design depends on unknown quantities. Certainly, to identify a design for optimal estimation of $\boldsymbol{\beta}$, the correct subset of active effects must be identified. Bayesian approaches that express uncertainty about the correct subset enable construction of optimality criteria that account for this uncertainty. Such approaches typically find a design that optimizes a criterion which is averaged over many possible subsets. DuMouchel and Jones (1994) exploited this idea with a formulation in which some effects have uncertainty associated with whether they are active. Meyer et al. (1996) extended the prior distributions of Box and Meyer (1993) and constructed a "model discrimination" design criterion. The criterion is based on a Kullback–Leibler measure of dissimilarity between

predictions from two competing models and it averages this dissimilarity over all possible pairs of models. Averaging is weighted according to the prior probability of the models, thus incorporating prior information into the design criterion. Bingham and Chipman (2002) used weak heredity, as defined in (22), in a similar criterion based on the Hellinger distance.

Even in seemingly straightforward cases, such as a 16-run design, use of prior information can lead to the selection of nonregular designs. For example, Bingham and Chipman (2002) found that, if sufficiently small values of $\pi$ were used when looking for a six-factor, two-level design in 16 runs, a nonregular fractional factorial design was optimal. Under their criterion, nonregular designs were chosen over regular fractional factorial designs because they are better at estimating models containing interactions as well as main effects. Regular fractional factorial designs enable estimation of many linear main effects, at the cost of estimability of interactions. Some large models containing only linear main effects may actually seem implausible. For example, prior distribution (22) with $\pi = 0.41$ puts over 600 times more prior probability on a model with effects *A, B, AC* than a model with linear effects *A, B, C, D, E, F*. This leads the design criterion to select designs that sacrifice simultaneous estimability of all linear effects for the ability to estimate more interactions.

## 7 Discussion

Although the presentation here has focused on linear models with Gaussian errors, similar ideas may be applied to subset selection in other models. For example, George et al. (1995) extended the nonconjugate $\boldsymbol{\beta}$ prior distribution (5) to a probit regression model for a binary response. They exploited a relationship between a probit regression and a linear regression with normal errors. Let $Y_i = \mathrm{x}_i'\beta + \epsilon$, with $\epsilon \sim N(0, 1)$. Instead of observing $Y_i$ we observe a binary response $Z_i$ that is a thresholding of $Y_i$. Thus, we observe $Z_i = 1$ if $Y_i > 0$ and $Z_i = 0$ otherwise. Then $\Pr(Z_i = 1) = \Pr(Y_i > 0) = \Phi(\mathrm{x}_i'\boldsymbol{\beta})$ with $\Phi$ being the standard normal cumulative probability function. This is the probit model for a binary response. George et al. (1995) treated $Y$ as missing data and use the data augmentation approach of Tanner and Wong (1987) to simulate the unobserved $Y$. The Gibbs sampler for (5) can be applied to $Y$, so the extended algorithm alternates between draws of $\boldsymbol{\beta}$ and the unobserved latent variable $Z$.

Other more general modifications and extensions are possible. If an MCMC sampler for a model exists, then an additional step incorporating a draw of the subset indicator $\boldsymbol{\delta}$ should be possible. For example, in a generalized linear model with regression coefficient vector $\boldsymbol{\beta}$, dispersion parameter $\phi$, and a fixed subset of active effects, an MCMC sampler might be available for the full joint posterior distribution $p(\boldsymbol{\beta}, \phi|\boldsymbol{Y})$. To generalize to subset selection, it is necessary to draw from $p(\boldsymbol{\beta}, \phi, \boldsymbol{\delta}|\boldsymbol{Y})$. The draws for $\phi$ are unchanged. The draws for $\boldsymbol{\beta}$ are the same, except that the prior variances will be determined by $\boldsymbol{\delta}$. The draw for $\boldsymbol{\delta}$ is carried out one element at a time using the conditional distribution

$p(\delta_j|\delta_1 \ldots, \delta_{j-1}, \delta_{j+1}, \ldots, \delta_h, \boldsymbol{\beta}, \phi, \boldsymbol{Y})$. This conditional probability distribution will be a Bernoulli draw, with the probability of $\delta_j = 1$ depending on the ratio of two densities of $\beta_j$ (one with $\delta_j = 0$, the other with $\delta_j = 1$). One paper that developed such a sampler (without subset selection) is that of Dellaportas and Smith (1993), in which a Gibbs sampling scheme was used for generalized linear models and proportional hazards models.

Analytic approaches in which the marginal posterior distribution of $\boldsymbol{\delta}$ is obtained by integrating the posterior distribution with respect to $\boldsymbol{\beta}$ and $\phi$, are also possible. Analytic approximations such as the Laplace approximation (Tierney and Kadane, 1986) are necessary to obtain a closed-form expression for the marginal posterior distribution $p(\boldsymbol{\delta}|\boldsymbol{Y})$.

Another interesting problem in which prior distributions in Bayesian subset selection might be used is in situations in which there is complete aliasing between effects. By adding information about relationships between effects in the form of heredity prior distributions, the posterior distribution can be used to disentangle the most likely effects. Chipman and Hamada (1996) discovered such a pattern, in which there is support for the model $A, C, E, H, AE = CH = BF = DG$. The last four effects are aliased. Two of these ($BF$, $DG$) are discarded automatically because they do not obey heredity. Two submodels involving the other two are identified: *C, E, H, CH* and *A, C, E, AE*, with the former providing better fit and consequently receiving higher posterior probability.

The emphasis of this chapter is very much on subset *selection*, because the goal of screening experiments is to identify factors that influence the response. An alternative technique is Bayesian model averaging in which predictions are averaged across all possible subsets (or a representative sample), using posterior probability as weights. A review of Bayesian model averaging is given in Hoeting et al. (1999). The consensus among researchers in the field seems to be that model averaging produces better prediction accuracy than selection of a single subset. However, in screening experiments, selection remains paramount.

Software implementing the methods described in this chapter are available at the author's website, http://ace.acadiau.ca/math/chipmanh.

## *References*

Bingham, D. and Chipman, H. (2002). Optimal designs for model selection. Technical Report 388, University of Michigan.

Box, G. E. P. and Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics*, **28**, 11–18.

Box, G. E. P. and Meyer, R. D. (1993). Finding the active factors in fractionated screening experiments. *Journal of Quality Technology*, **25**, 94–105.

Chipman, H. (1996). Bayesian variable selection with related predictors. *The Canadian Journal of Statistics*, **24**, 17–36.

Chipman, H. A. (1998). Fast model search for designed experiments. In *Quality Improvement Through Statistical Methods*. Editor: B. Abraham, pages 205–220. Birkhauser, Boston.

Chipman, H. A. and Hamada, M. S. (1996). Comment on "Follow-up designs to resolve confounding in multifactor experiments." *Technometrics*, **38**, 317–320.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. In *Model Selection*. Editor: P. Lahiri, pages 65–116. Volume 38 of *IMS Lecture Notes—Monograph Series*, Institute of Mathematical Statistics, Beachwood.

Chipman, H., Hamada, M., and Wu, C. F. J. (1997). A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics*, **39**, 372–381.

Dean, A. M. and Voss, D. T. (1999). *Design and Analysis of Experiments*. Springer, New York.

Dellaportas, P. and Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, **42**, 443–459.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, third edition. John Wiley and Sons, New York.

DuMouchel, W. and Jones, B. (1994). A simple Bayesian modification of *D*-optimal designs to reduce dependence on an assumed model. *Technometrics*, **36**, 37–47.

Furnival, G. M. and Wilson, Robert W. J. (1974). Regression by leaps and bounds. *Technometrics*, **16**, 499–511.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. Chapman & Hall/CRC, Boca Raton, FL.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–374.

George, E. I., McCulloch, R. E., and Tsay, R. S. (1995). Two approaches to Bayesian model selection with applications. In *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*. Editors: D. A. Berry, K. A. Chaloner, and J. K. Geweke, pages 339–348. John Wiley and Sons, New York.

Hamada, M. and Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, **24**, 130–137.

Henkin, E. (1986). The reduction of variability of blood glucose levels. In *Fourth Supplier Symposium on Taguchi Methods*, pages 758–785. American Supplier Institute, Dearborn, MI.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, **14**, 382–417. (Correction: **15**, 193–195; corrected version available at www.stat.washington.edu/www/research/online/hoeting1999.pdf).

Lee, P. M. (2004). *Bayesian Statistics: An Introduction*, third edition. Arnold, London.

Meyer, R. D., Steinberg, D. M., and Box, G. (1996). Follow-up designs to resolve confounding in multifactor experiments (with discussion). *Technometrics*, **38**, 303–322.

Nelder, J. A. (1998). The selection of terms in response-surface models: How strong is the weak-heredity principle? *The American Statistician*, **52**, 315–318.

O'Hagan, A. and Forster, J. J. (2004). *Kendall's Advanced Theory of Statistics, volume 2B: Bayesian Inference*, second edition. Arnold, London.

Peixoto, J. L. (1990). A property of well-formulated polynomial regression models. *The American Statistician*, **44**, 26–30. (Correction: **45**, 82.)

R Development Core Team (2004). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**, 179–191.

Tanner, M. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–550.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Zellner, A. (1987). *An Introduction to Bayesian Inference in Econometrics*. Krieger, Melbourne, Florida.

# 12
# Analysis of Orthogonal Saturated Designs

Daniel T. Voss and Weizhen Wang

This chapter provides a review of special methods for analyzing data from screening experiments conducted using regular fractional factorial designs. The methods considered are robust to the presence of multiple nonzero effects. Of special interest are methods that try to adapt effectively to the unknown number of nonzero effects. Emphasis is on the development of adaptive methods of analysis of orthogonal saturated designs that rigorously control Type I error rates of tests or confidence levels of confidence intervals under standard linear model assumptions. The robust, adaptive method of Lenth (1989) is used to illustrate the basic problem. Then nonadaptive and adaptive robust methods of testing and confidence interval estimation known to control error rates are introduced and illustrated. Although the focus is on Type I error rates and orthogonal saturated designs, Type II error rates, nonorthogonal designs, and supersaturated designs are also discussed briefly.

## 1 Introduction

In the design and analysis of experiments in industry, screening, plays an important role in the early phases of experimentation. In Chapter 1, Montgomery and Jennings provide an overview of screening experiments and also give an introduction to regular fractional factorial designs which are often used in this context. Building upon this foundation, we give further consideration to the analysis of data collected using such designs. In particular, we describe methods that are appropriate when the design of the experiment produces just enough observations to allow estimation of the main effects and interactions of interest; that is, the design is *saturated*. We concentrate on methods of analysis that are *adaptive* in the sense that the estimator of the error variance is altered depending on the values of the estimated main effects and interactions from the experiment.

For motivation and illustration, we consider the plasma etching experiment discussed by Montgomery and Jennings in Chapter 1. The experimental design was a $2^{6-2}_{\text{IV}}$ regular fractional factorial design with 16 observations that allows the estimation of 15 factorial effects. The data and aliasing scheme are given in Tables 3 and 4 of Chapter 1.

The linear model, in matrix form, that we use for data analysis is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where the vector $\boldsymbol{Y}$ holds the response variables $Y_1, \ldots, Y_n$; the vector $\epsilon$ holds the error variables $\epsilon_1, \ldots, \epsilon_n$ and these are independent and normally distributed with constant variance $\sigma^2$; the vector $\boldsymbol{\beta}$ holds the unknown parameters $\beta_0, \beta_1, \ldots, \beta_h$; and $\boldsymbol{X}$ is the model matrix which is formulated as described below.

The least squares estimate of each main effect is the average of the eight observations at the high level of the factor minus the average of the eight observations at the low level and, likewise, the least squares estimate of each interaction effect is the average of an appropriate set of eight observations minus the average of the other eight observations, as described in Chapter 1, Section 2. In order for the parameters in our model to measure the main effects and interactions directly, the columns of the model matrix $\boldsymbol{X}$ are formed as follows. The first column consists of a column of ones corresponding to the intercept parameter $\beta_0$. For each main effect parameter, the elements of the corresponding column of $\boldsymbol{X}$ consist of $+0.5$ for the high level of the factor and $-0.5$ for the low level. For the parameter measuring the interaction between factors $i$ and $j$, the entries in the corresponding column of $\boldsymbol{X}$ are obtained by multiplying the elements of the $i$ and $j$ main effects columns, and then multiplying by 2 so that all elements are again $+0.5$ or $-0.5$.

When an experiment is carefully conducted and the correct model is used, then independence of the response variables is often a reasonable assumption. The experimenters of the plasma etching investigation were reportedly comfortable with the assumptions of normality and constant variance, based on their prior experience with similar experiments. From the 16 observations collected using this design, under the above model assumptions, there are 15 independent factorial effect estimators that can be used for the analysis. We denote these estimators by $\hat{\beta}_i, i = 1, 2, \ldots, 15$, corresponding to the effects $A, B, C, D, E, F, AB, AD, AE, AF, BD, BE, BF, ABD, ABF$. These values of $i$ include a representative effect from each set of aliases for this design; see Chapter 1 for a discussion of aliases and see Table 4 of that chapter for the defining relation and aliases for the plasma etching experiment under consideration here.

Under the assumptions of model (1), the least squares estimators of the 15 main effects and interactions are independently normally distributed with equal variances, and each estimator provides an unbiased estimate of the corresponding effect. This effect is a factorial treatment contrast together with its aliases (as explained in Chapter 1). Independence of the estimators of these effects is a consequence of using an *orthogonal design*, because an orthogonal design is one that yields uncorrelated estimators under the assumptions of model (1) and uncorrelated estimators are independent under normality. For the plasma etching experiment, the least squares estimates are given in Table 1 of this chapter and are used to illustrate various methods of analysis.

TABLE 1. Factorial effect least squares estimates and squared estimates for the plasma etching experiment of Chapter 1

| Effect | $\hat{\beta}_i$ | $\hat{\beta}_i^2$ |
|---|---|---|
| A | −175.50 | 30800.25 |
| AB | 106.75 | 11395.56 |
| E | 103.50 | 10712.25 |
| B | 58.00 | 3364.00 |
| BE | −53.75 | 2889.06 |
| ABF | −29.75 | 885.06 |
| AE | 27.25 | 742.56 |
| D | 18.75 | 351.56 |
| F | −18.75 | 351.56 |
| C | −18.50 | 342.25 |
| BF | −16.00 | 256.00 |
| AF | −13.00 | 169.00 |
| ABD | −5.75 | 33.06 |
| AD | 4.50 | 20.25 |
| BD | 3.00 | 9.00 |

Given a regular fraction of a $2^f$ experiment and independent response variables, the estimators described above have constant variance even if the individual response variables do not. Also, the estimators are approximately normally distributed by the Central Limit Theorem. However, if the response variables have unequal variances, this unfortunately causes the estimators to be correlated and, therefore, dependent. The use of the data analysis to assess whether the levels of some factors affect the response variability is, itself, a problem of great interest due to its role in robust product design; see Chapter 2 for a discussion of available methods and analysis. In the present chapter, we consider situations in which the estimators are independent.

One additional premise, fundamental to the analysis of data from screening experiments, is the assumption of *effect sparsity*, namely, that very few of the effects under study are sizable. In a screening experiment, it is common for an investigator to study as many factors as possible in the experiment, but there is usually a restriction on the number of observations that can be run. As a result, screening experiments often include no replication and so provide no pure estimate of error variance. Furthermore, such experiments are often designed to be saturated (having just enough observations to estimate all of the effects, but leaving no degrees of freedom for error). Thus, there is no mean squared error with which to estimate the error variance independently of effect estimates. The lack of an independent estimator of variance means that standard methods of analysis, such as the analysis of variance and confidence intervals and tests based on the *t*-distribution, do not apply. Nonetheless, provided that effect sparsity holds, the use of a saturated design often leads to the estimates of the large effects standing out relative to the others; this is fundamental to the effective analysis of data from a saturated design.

A traditional approach to the analysis of data from an orthogonal saturated design utilizes half-normal plots of the effect estimates. This approach was introduced by Daniel (1959) and is illustrated in Figure 3 of Chapter 1 for the plasma etching data. In a half-normal plot, the ordered absolute values of the estimates are plotted against their expected values, or half-normal scores, under the assumption that the estimators all have mean zero in addition to being normally distributed with equal variances. If only a few estimates are relatively large in absolute value, then the corresponding points tend to stand out in the plot away from a line through the points corresponding to the absolute values of the smaller estimates. Daniel advocated using this approach *iteratively*: if the largest estimate is determined, often subjectively, to correspond to a nonzero effect, then that estimate is removed before the next step of the analysis; at the next step, the half-normal plot is regenerated using the remaining effects; the largest remaining estimate is then evaluated. This process is iterated until the largest remaining estimate does not stand out. In practice, such iteration is seldom done. Without iteration, a reasonable interpretation of the half normal plot in Figure 3 of Chapter 1 is that five effects stand out, these being the effects $A$, $AB$, $E$, $B$, and $BE$, or their aliases (for example, $BE = AC$).

From an historical perspective, the analysis of orthogonal saturated designs was considered initially by Birnbaum (1959) and Daniel (1959). In addition to half-normal plots for the subjective analysis of orthogonal saturated designs, Daniel (1959) also considered more formal, objective methods of analysis of such designs, as did Birnbaum (1959) in a companion paper. Each considered testing for a nonzero effect amongst $h$ effects, assuming that at most one effect is nonzero. Birnbaum provided a most powerful test of this simple hypothesis (see Section 2 for the definition of a most powerful test). His test is based on the proportion of the total variation that is explained by the largest estimated effect. Such a test could be iterated to test for more than one nonzero effect, but then true error rates and the choice of a best test become unclear. Birnbaum also sought optimal rules for deciding which, and how many, effects are nonzero when at most two effects are truely nonzero and noted that the problem was then already quite complex.

Subsequently, Zahn (1969, 1975ab) considered some variations on the iterative methods of Daniel (1959) and Birnbaum (1959), but his results were primarily empirical. The subjective use of half-normal plots remains a standard methodology for the analysis of orthogonal saturated designs, but the development of objective methods is progressing rapidly.

Box and Meyer (1986, 1993) provided Bayesian methods for obtaining posterior probabilities that effects are *active*; see Chapter 11, Section 2, for more details. There followed a flurry of papers proposing new frequentist methods, giving refinements of the methods, and making empirical comparisons of the many variations. Hamada and Balakrishnan (1998) provided an extensive review of these methods, including a Monte Carlo-based comparison of the "operating characteristics" of the methods; that is, a comparison of the power of the methods for a variety of combinations of effect values (*parameter configurations*). They found that comparison of methods is difficult for various reasons. For example, some

are intended for individual inferences and others for simultaneous inference. Each method is designed to be "robust", as discussed in Section 3, but each method has its own inherent "breakdown point" with respect to the number of nonnegligible effects that can be identified. Still, there are commonalities of the better methods. The preferred methods typically use the smaller ordered absolute values of the estimated effects or the corresponding sums of squares to obtain a robust estimate of variability. We refer the reader to Hamada and Balakrishnan (1998) for further details on the many proposed methods.

The most influential of these methods is a "quick and easy" method introduced by Lenth (1989). Lenth's method was ground breaking because, in addition to being simple, it is robust to the presence of more than one large effect and it is adaptive to the number of large effects. Furthermore, empirical studies suggest that it maintains good power over a variety of parameter configurations. We use Lenth's method to illustrate the concepts of "robust" and "adaptive" in Section 3 and apply it to the plasma etching study of Chapter 1. In Section 4, we introduce robust methods of confidence interval estimation, some adaptive and some not, but all known to provide at least the specified confidence level under all parameter configurations. Section 5 contains some analogous results for hypothesis testing. These confidence interval and testing methods are again illustrated using the data from the plasma etching study. The chapter concludes with a broader discussion of the issues for the analysis of unreplicated factorial designs. First though, in Section 2, we give our formulation of the factor screening problem.

## 2 Formulation of the Factor Screening Problem

For a given experiment with $f$ factors conducted using a $2^{f-q}$ fractional factorial design, there are $h = 2^{f-q} - 1$ independent factorial effect estimators $\hat{\beta}_i, i = 1, \ldots, h$, where $\hat{\beta}_i \sim N(\beta_i, \sigma^2_{\hat{\beta}})$. Suppose that effect sparsity holds so that most of the effects $\beta_i$ are zero (or negligible), with only a few of the effects being large in magnitude.

The basic factor screening problem is to determine which effects are large and which are small. Any factor found to have only negligible main effects and interactions requires no further investigation and so need not be considered in subsequent experimentation. The primary objective of a screening experiment is, therefore, to screen out unimportant factors so that subsequent experiments can focus on studying the important factors without being unduly large. Also, any factors found to have large effects deserve further study, so their identification is obviously of value.

In the language of hypothesis testing, it is generally agreed that Type I and Type II errors are both of importance in screening experiments. If a Type I error is made, the consequence is that an unimportant factor remains under investigation for one or more subsequent rounds of experimentation, using additional resources and perhaps slowing progress. On the other hand, if a Type II error is made, an important factor may be excluded from future experiments and this could undermine the

success of the entire study. In view of this, one could argue that confidence intervals are more useful than hypothesis testing, provided that the confidence intervals are tight enough to pin down adequately the magnitude of each effect. Perhaps rejecting, or failing to reject, the null hypothesis that an effect is zero in the absence of power calculations may be of little value. Or, one might argue that a completely different formulation of the problem is needed, perhaps along the lines of bioequivalence testing (see Brown et al., 1997), where the goal is to demonstrate similarity of effects rather than differences. For example, one could try to demonstrate that certain effects are close to zero so that the corresponding factors merit no further consideration.

The above discussion suggests possible difficulties in formulating the problem of data analysis for screening experiments. However, it is perhaps more accurate to say that, even if Type II errors are as, or more, important than Type I errors in the analysis of screening experiments, it is more difficult to deal with Type II errors. (This is because the probability of making Type II errors depends on the parameter configuration under composite alternative hypotheses.) We can obviously avoid making Type II errors by always asserting that all effects are nonzero, so never screening out any factors, but then the primary goal of a screening experiment cannot be achieved. So, in searching for the methods that are best at detecting large effects in the analysis of screening experiments, one must strike a balance between Type I and Type II errors.

To be able to compare methods even-handedly, we have chosen to rely on a fundamental approach in statistical hypothesis testing; namely, we seek the most powerful level-$\alpha$ test. A test is of *level-*$\alpha$ if the supremum of the probability of making a Type I error over the null hypothesis space is at most $\alpha$. If this supremum is equal to $\alpha$, the test is said to be of *size* $\alpha$. Amongst level-$\alpha$ tests, one is a *most powerful test* if, loosely speaking, it is the most likely to detect nonzero effects, large or small. In the analysis of saturated designs, establishing that a test is level-$\alpha$ is complicated by the fact that all of the parameters are of interest yet, in testing any single effect, the other effects are nuisance parameters. For the methods of analysis proposed in the literature, critical values are invariably determined under the assumption that all of the effects $\beta_i$ are zero; we call this the *null case*. However, for most of the available methods of data analysis, it still remains an open problem to prove that use of critical values obtained under the null distribution yields a level-$\alpha$ test so that the Type I error rate cannot exceed $\alpha$. If the answer to this open problem is "yes," then use of these critical values would yield size-$\alpha$ tests. Such may well be the case for orthogonal designs but, most curiously, Wang and Voss (2001b) provided a counterexample for a particular method of analysis of nonorthogonal saturated designs.

In view of these considerations, a reasonable goal is to seek methods of analysis for screening experiments that include powerful tests of specified size or level and exact confidence intervals that are tight. An *exact confidence interval* is analogous to a test of specified size—"exact" means that the confidence level is at least as high as the level claimed and the level claimed is the greatest lower bound on the confidence level.

## 3 Robust Adaptive Methods

The most influential method of analysis of orthogonal saturated designs yet proposed is the robust adaptive method of Lenth (1989). The "quick and easy" method that he proposed is based on the following estimator of the standard deviation, $\sigma_\beta$, of the effect estimators $\hat{\beta}_i$. This estimator is "robust" and "adaptive", concepts which are explained in detail after the following description of the method.

First, obtain an initial estimate of $\sigma_\beta$:

$$\hat{\sigma}_o = 1.5 \times (\text{the median of the absolute estimates } |\hat{\beta}_i|). \tag{2}$$

For the plasma etching experiment, the median absolute estimate is 18.75, which can be found as the eighth listed value in Table 1. This yields $\hat{\sigma}_o = 28.13$ from (2). If the estimators $\hat{\beta}_i$ are all normally distributed with mean zero and common standard deviation $\sigma_\beta$, then $\hat{\sigma}_o$ is approximately unbiased for $\sigma_\beta$.

Second, calculate an updated estimate

$$\hat{\sigma}_L = 1.5 \times (\text{the median of those } |\hat{\beta}_i| \text{ that are less than } 2.5\hat{\sigma}_o),$$

where subscript $L$ denotes Lenth's method. For the plasma etching experiment, $2.5\hat{\sigma}_o = 70.31$. Each of the three largest absolute estimates exceeds this value and so is set aside. The median of the remaining 12 absolute estimates is 18.63. Thus $\hat{\sigma}_L = 27.94$, which is slightly smaller than $\hat{\sigma}_o$.

Lenth (1989) referred to this estimator $\hat{\sigma}_L$ as the *pseudo standard error* of the estimators $\hat{\beta}_i$. He recommended using the test statistic $\hat{\beta}_i/\hat{\sigma}_L$ to test the null hypothesis $\beta_i = 0$ and recommended using the quantity $(\hat{\beta}_i - \beta_i)/\hat{\sigma}_L$ to construct a confidence interval for $\beta_i$.

Critical values for individual tests and confidence intervals are based on the *null distribution* of $|\hat{\beta}_i|/\hat{\sigma}_L$, that is, on the distribution of this statistic when all effects $\beta_i$ are zero. Lenth proposed a $t$-distribution approximation to the null distribution, whereas Ye and Hamada (2000) obtained exact critical values by simulation of $|\hat{\beta}_i|/\hat{\sigma}_L$ under the null distribution. From their tables of exact critical values, the upper 0.05 quantile of the null distribution of $|\hat{\beta}_i|/\hat{\sigma}_L$ is $C_L = 2.156$. On applying Lenth's method for the plasma etching experiment and using $\alpha = 0.05$ for individual inferences, the minimum significant difference for each estimate is calculated to be $c_L \times \hat{\sigma}_L = 60.24$. Hence, the effects $A$, $AB$, and $E$ are declared to be nonzero, based on individual 95% confidence intervals.

From empirical comparisons of various proposed methods of analysis of orthogonal saturated designs (Hamada and Balakrishnan, 1998; Wang and Voss, 2003), Lenth's method can be shown to have competitive power over a variety of parameter configurations. It remains an open problem to prove that the null case is the least favourable parameter configuration.

We now discuss what it means for a method of analysis to be "robust" or "adaptive". Lenth's method is *adaptive* because of the two-stage procedure used to obtain the pseudo standard error. The pseudo standard error $\hat{\sigma}_L$ is computed from

the median of most of the absolute estimates, but how many are excluded from the calculation depends on the data. In other words, the procedure adapts itself to the data, and it attempts to do so efficiently. It seems reasonable to believe that $\beta_i \neq 0$ if $|\hat{\beta}_i| > 2.5\hat{\sigma}_o$ because, for a random variable $X$ having a normal distribution with mean zero and standard deviation $\sigma$, $P(|X| > 2.5\sigma) \approx 0.0124$. In a sense, one is pre-testing each hypothesis

$$H_{0,i} : \beta_i = 0$$

in order to set large estimates aside in obtaining the pseudo standard error, which is then used for inference on the remaining effects $\beta_i$.

Consider now robustness. If the estimators $\hat{\beta}_i$ are computed from independent response variables then, as noted in Section 1, the estimators have equal variances and are usually at least approximately normal. Thus the usual assumptions, that estimators are normally distributed with equal variances, are approximately valid and we say that there is inherent *robustness* to these assumptions. However, the notion of *robust methods* of analysis for orthogonal saturated designs refers to something more. When making inferences about any effect $\beta_i$, all of the other effects $\beta_k$ $(k \neq i)$ are regarded as nuisance parameters and "robust" means that the inference procedures work well, even when several of the effects $\beta_k$ are large in absolute value. Lenth's method is robust because the pseudo standard error is based on the median absolute estimate and hence is not affected by a few large absolute effect estimates. The method would still be robust even if one used the initial estimate $\hat{\sigma}_o$ of $\sigma_\beta$, rather than the adaptive estimator $\hat{\sigma}_L$, for the same reason.

Any robust method has a *breakdown point*, which is the percentage of large effects that would make the method ineffective. For Lenth's method, if half or more of the effect estimates are very large in magnitude, then $\hat{\sigma}_o$ will be large and hence so will $\hat{\sigma}_L$, causing the method to lose so much power that the method breaks down. Hence, the breakdown point is about 50%. One could lower the breakdown point by using, for example, the 30th percentile of the absolute estimates rather than the median to estimate $\sigma_\beta$. However, this would increase the variability of the pseudo standard error, which would reduce power when there truly is effect sparsity.

In summary, Lenth's method is robust in the sense that it maintains good power as long as there is effect sparsity and it is adaptive to the degree of effect sparsity, using a pseudo standard error that attempts to involve only the estimates of negligible effects.

Like many methods of analysis of orthogonal saturated designs proposed in the literature, the critical values for Lenth's method are obtained in the null case (all $\beta_i$ zero), assuming this is sufficient to control the Type I error rates. This raises the question: can one establish analytically that Lenth's and other proposed methods do indeed provide the claimed level of confidence or significance under standard model assumptions? The rest of this chapter concerns methods for which the answer is "yes."

# 4 Robust Exact Confidence Intervals

In this section, we discuss the construction of individual confidence intervals for each factorial effect $\beta_i$, based only on the least squares estimates $\hat{\beta}_1, \ldots, \hat{\beta}_h$. An exact 100 $(1-\alpha)$% confidence interval for $\beta_i$ is analogous to a size-$\alpha$ test of the null hypothesis $\beta_i = 0$, against a two-sided alternative. In testing this hypothesis, the probability of making a Type I error depends on the values of the other parameters $\beta_k, k \neq i$. Such a test would be of size $\alpha$, for example, if under the null hypothesis, the probability of making a Type I error were exactly $\alpha$ when $\beta_k = 0$ for all $k \neq i$ and at most $\alpha$ for any other values of the $\beta_k$ for $k \neq i$. By analogy, a confidence interval for $\beta_i$ would be an exact $100(1-\alpha)$% confidence interval if the confidence level were exactly 100 $(1-\alpha)$% when $\beta_k = 0$ for all $k \neq i$ and at least 100 $(1-\alpha)$% for any values of the $\beta_k$ for $k \neq i$. Inference procedures that control the error rates under all possible parameter configurations are said to provide strong control of error rates, see Hochberg and Tamhane (1987) and also Chapter 6. For a confidence interval, the error rate is at most $\alpha$ if the confidence level is at least 100 $(1-\alpha)$%.

When screening experiments are used, it is generally anticipated that several effects may be nonzero. Hence, one ought to use statistical procedures that are known to provide strong control of error rates. It is not enough to control error rates only under the complete null distribution. This section discusses exact confidence intervals. Size-$\alpha$ tests are considered in Section 5.

## *4.1 Non-Adaptive Confidence Intervals*

The first confidence interval for the analysis of orthogonal saturated designs that provided strong control of the error rate was established by Voss (1999). His confidence interval for $\beta_i$ excludes $\hat{\beta}_i$ from the computation of the standard error and is obtained using the random variable

$$(\hat{\beta}_i - \beta_i)/\hat{\sigma}_V,$$

where the denominator is the square root of

$$\hat{\sigma}_V^2 = \frac{\sum_{k=1}^{u} \hat{\beta}_{(k)}^2}{u}, \tag{3}$$

which is the mean squared value of the $u$ smallest of the $h-1$ effect estimates excluding $\hat{\beta}_i$, and where $u$ is specified before the data are examined. Here $\hat{\beta}_{(k)}^2$ denotes the $k$th smallest of the $h-1$ squared estimates $\hat{\beta}_k^2$ for $k \neq i$.

Let $c_V$ be the upper-$\alpha$ critical value, obtained as the upper-$\alpha$ quantile of the distribution of $|\hat{\beta}_i|/\hat{\sigma}_V$ when all effects are zero. Voss (1999) showed that

$$\hat{\beta}_i \pm c_V \hat{\sigma}_V$$

is an exact 100 $(1-\alpha)$% confidence interval for $\beta_i$. This result was obtained from a basic but obscure Stochastic Ordering Lemma, which says that

$$|\hat{\beta}_i - \beta_i|/\hat{\sigma}_V \tag{4}$$

is stochastically largest under the complete null distribution. This follows because (4) is a nonincreasing function of $\hat{\beta}_k^2$ for each $k \neq i$, the estimators $\hat{\beta}_k$ $(k = 1, \ldots, h)$ are independent, and the distribution of each $\hat{\beta}_k^2$ $(k \neq i)$ is increasing in $\hat{\beta}_k^2$. As a consequence, if

$$P(|\hat{\beta}_i - \beta_i|/\hat{\sigma}_V \leq c_V) = 1 - \alpha$$

under the null distribution, then

$$P(|\hat{\beta}_i - \beta_i|/\hat{\sigma}_V \leq c_V) \geq 1 - \alpha$$

under any parameter configuration. This stochastic ordering result was obtained independently by Alam and Rizvi (1966) and Mahamunulu (1967).

We now apply Voss' method to the data from the plasma etching experiment to construct individual 95% confidence intervals for each effect using $u = 8$. The pooling of 8 sums of squares into the denominator (3) provides a reasonably robust procedure without undue loss of power. One could, of course, pool more than 8 sums of squares into the denominator, as one would usually anticipate greater effect sparsity—more than 8 negligible effects—in a screening experiment. Still, 8 provides a reasonable trade-off between power and robustness. Also, for simultaneous confidence intervals, Dean and Voss (1999) provided critical values for this choice because, in a single replicate $2^4$ factorial experiment, an inactive factor is involved in 8 inactive main effects and interactions which could then be used to provide the denominator (3) in Voss' method.

On application of Voss' method for $h = 15$ total effects, $u = 8$ smallest effects, and 95% individual confidence level, we find by simulation that the critical value is $c_V = 5.084$. This and subsequent simulated critical values in this chapter are obtained by generating a large number of sets of estimates under the null distribution (that is, with mean zero and standard deviation one), computing the relevant statistic for each set of estimates, and using the upper $\alpha$ quantile of the resulting empirical distribution of values of the statistic as the critical value. Sample programs for computing critical values are available at http://www.wright.edu/~dan.voss/screening.htm.

For any of the seven largest estimates, expression (3) gives $\hat{\sigma}_V^2 = 191.59$, so the minimum significant difference is $c_V\hat{\sigma}_V = 70.37$. From Table 1, we see that three effects have estimates larger than 70.37 in absolute value, namely, *A, AB*, and *E*. Hence, individual 95% confidence intervals for these three effects do not include zero, so these effects are declared to be nonzero. No other effects can be declared to be nonzero using this method. These results match those obtained using Lenth's individual 95% confidence intervals.

Voss and Wang (1999) showed that simultaneous confidence intervals could be obtained using a similar, but more complicated, technical justification. For

simultaneous intervals, the computation of the critical value is based on the null distribution of the maximum of the $h$ random variables $|\hat{\beta}_i|/\hat{\sigma}_V$ and the upper $\alpha$ quantiles can be obtained via simulation, as described above. The critical values are provided in Table A.11 of Dean and Voss (1999). We do not illustrate this method here but instead illustrate adaptive simultaneous confidence intervals in the following section. Based on power simulations conducted by Wang and Voss (2003), adaptive methods appear to have better minimax power and competitive average power when compared to nonadaptive methods over a variety of parameter configurations.

## *4.2 Adaptive Confidence Intervals*

We now extend the above ideas to obtain adaptive individual confidence intervals for each effect $\beta_i$ and again apply the methods to the plasma etching example. In developing such intervals that strongly control the error rate, the motivation of Wang and Voss (2001a, 2003) was the approach used when examining the half-normal probability plot. This involves looking for a jump in the magnitude of the absolute estimates, or their squared values, in order to determine how many of these should be pooled into the estimate of $\sigma_\beta^2$. Below, we describe the methodology of Wang and Voss (2003) and, for simplicity, we concentrate on the special case of their general theory that is most useful in practice.

Suppose that one allows the possibility of pooling into the estimate of error the $j$ smallest of the $h - 1$ squared estimates, excluding $\hat{\beta}_i^2$, for some prespecified set $J$ of choices for $j$. For example, for $h = 15$ effects, one might consider pooling either 8 or 12 of the 14 available squared estimates, because 12 might give very good power if only one or two effects are nonnegligible, but 8 would be a better choice if there happens to be less effect sparsity. This corresponds to taking $J = \{8, 12\}$, and simulations by Wang and Voss (2003) found that this choice provides good power under a variety of parameter configurations. Let

$$\hat{\sigma}_j^2 = w_j \sum_{k=1}^{j} \hat{\beta}_{(k)}^2/j \tag{5}$$

denote the mean squared value of the $j$ smallest squared estimates (excluding $\hat{\beta}_i$), scaled by a prespecified weight $w_j$, and let $\hat{\sigma}_{\min}^2$ be the minimum value of all the $\hat{\sigma}_j^2$ for values of $j$ in the prechosen set $J$; that is,

$$\hat{\sigma}_{\min}^2 = \min\{\hat{\sigma}_j^2 | j \in J\}. \tag{6}$$

Because the $\hat{\beta}_{(k)}^2$ are ordered in increasing value, $\sum_{k=1}^{j} \hat{\beta}_{(k)}^2/j$ is increasing in $j$. So, the condition that $w_j < w_{j'}$ for $j > j'$ is imposed on the constants $w_j$ in order for $\hat{\sigma}_{\min}$ to be adaptive, namely, so that any $j$ could yield the minimum value of $\hat{\sigma}_j^2$. Also, $\hat{\sigma}_{\min}$ is a nondecreasing function of each $\hat{\beta}_k^2$ for $k \neq i$, providing the means to establish strong control of error rates via the Stochastic Ordering Lemma.

An application of this lemma shows that an exact $100\,(1-\alpha)\%$ confidence interval for $\beta_i$ is given by

$$\hat{\beta}_i \pm c_{\min}\hat{\sigma}_{\min},$$

where $c_{\min}$ denotes the upper-$\alpha$ quantile of the null distribution of $|\hat{\beta}_i|/\hat{\sigma}_{\min}$.

Some further guidance is needed concerning specification of the set $J$ and the weights $w_j$ for $j \in J$. When exactly $j$ of the effects are zero or negligible, it is desirable that $\hat{\sigma}_{\min}$ is equal to $\hat{\sigma}_j$ and the chance of this happening is greater for smaller $w_j$. This provides some basis for choosing the $w_j$ using any existing knowledge concerning the likely number of negligible effects. However, one generally does not know how many effects are negligible; hence the desire for a robust adaptive method. Wang and Voss (2003) conducted an empirical power study of the procedure for various choices of $J$ and $w_j$ for $j \in J$ for the analysis of 15 effects. A choice of $J$ that yielded good minimax and average power over a variety of parameter configurations was the set $J = \{8, 12\}$, for which either the 8 or 12 smallest squared estimates are pooled to estimate the variance. Furthermore, for this choice, each weight $w_j$, for $j \in \{8, 12\}$, was chosen to make $\hat{\sigma}_j^2$ an unbiased estimator of the common variance of the estimators $\hat{\beta}_i$ when all effects $\beta_i$ are zero. To apply this method, each value $w_j$ can be obtained by simulation by computing the average value of $\sum_{k=1}^{j} Z_{(k)}^2/j$—that is, the average of the $j$ smallest squared values of $h-1$ pseudo standard normal random variables—and then taking the reciprocal of this average.

If we apply this method to the plasma etching experiment and compute individual 95% confidence intervals using $J = \{8, 12\}$, we obtain by simulation the values $w_8 = 4.308$, $w_{12} = 1.714$, and $c_{\min} = 2.505$. For the effects *A, AB*, and *E* corresponding to the three largest estimates, $\hat{\sigma}_8^2 = 825.35$ and $\hat{\sigma}_{12}^2 = 1344.54$, so $\hat{\sigma}_{\min}^2 = 825.35$ and the minimum significant difference is $c_{\min}\hat{\sigma}_{\min} = 71.97$. Therefore the effects *A, AB*, and *E* are declared to be significantly different from zero. The effects with the next largest estimates are *B* and *AC*. These are not significant based on the same minimum significant difference. We note, in passing, that the value of $\hat{\sigma}_{12}^2$ is larger for these effects, because it is computed using the 12 smallest squared estimates apart from the one for which the confidence interval is being constructed.

Wang and Voss (2003) showed that simultaneous confidence intervals could be obtained in a similar way, by computing the critical value based on the null distribution of the maximum of the $h$ random variables $|\hat{\beta}_i|/\hat{\sigma}_{\min}, i = 1, \ldots, h$, where, for each $i$, $\hat{\sigma}_{\min}$ is a function of the estimators excluding $\hat{\beta}_i$. To obtain simultaneous 95% confidence intervals for all 15 effects, the simulated critical value provided by Wang and Voss (2003) is $c_{\min} = 6.164$. For examining each of the 7 largest estimates, $\hat{\sigma}_{\min}^2$ is again equal to $\hat{\sigma}_8^2 = 825.35$ from (5) and (6). So we find that the minimum significant difference for simultaneous 95% confidence intervals is

$$c_{\min}\hat{\sigma}_{\min} = 177.08.$$

The largest effect estimate, $\hat{\beta}_A = -175.50$, has magnitude just under this more stringent threshold for simultaneous inference and thus none of the effects are found to differ significantly from zero.

Obviously, simultaneous 95% confidence intervals are more conservative than individual 95% confidence intervals, explaining the lack of significant results in this case. We advocate that both individual and simultaneous confidence intervals be used, because they provide different information and are both useful. The finding that three effects are significant using individual, but not simultaneous, 95% confidence intervals suggests the possibility of false positives, for example, whereas the significant results would be more believable if the simultaneous confidence intervals identified the same effects as being significant.

Simulations of Wang and Voss (2003) show little difference in power between their adaptive confidence intervals using $J = \{8, 12\}$ and the confidence intervals of Lenth (1989), though the former have the advantage that control of Type I error rates is established.

## 5 Robust Size-$\alpha$ Tests

In this section we focus on hypothesis testing and discuss both individual and simultaneous tests for detecting nonzero effects. Of special interest are the step-down tests described in Section 5.2, as these offer improved power over single-step tests.

### *5.1 Individual and Simultaneous Single-Step Tests*

Adaptive, robust single-step tests of size $\alpha$, both individual and simultaneous, can be based on the corresponding confidence intervals already discussed. To test the hypothesis

$$H_{0,i} : \beta_i = 0$$

for each fixed $i$, or to test these hypotheses simultaneously, one may simply check whether the corresponding Wang and Voss (2003) individual or simultaneous confidence intervals include zero. The test procedure is: reject each null hypothesis $H_{0,i}$ if and only if the confidence interval for $\beta_i$ excludes zero. This testing procedure controls the error rate and uses the data adaptively.

Better yet, one can obtain adaptive robust tests that are more easily implemented and are still of the specified size. For testing each null hypothesis $H_{0,i} : \beta_i = 0$, one need not exclude the corresponding estimator $\hat{\beta}_i$ from the computation of the denominator or standard error, because $\beta_i = 0$ under the null hypothesis.

For example, to obtain an individual test of $H_{0,i} : \beta_i = 0$, Berk and Picard (1991) proposed rejecting the null hypotheses for large values of the test statistic

$$\frac{\hat{\beta}_i^2}{\sum_{k=1}^{u} |\hat{\beta}|_{(k)}^2 / u},$$

where the denominator is the mean value of the $u$ smallest squared estimates computed from all $h$ estimates, for a pre-specified integer $u$. This test controls the error rate because the test statistic is nonincreasing in $\hat{\beta}_k^2$ for each $k \neq i$. Also, because the denominator is the same for testing each hypothesis $H_{0,i}$, implementation of the test is simple relative to implementation of corresponding confidence interval procedures.

Analogously, Voss and Wang (2005) proposed individual and simultaneous adaptive tests based on $\hat{\beta}_i / \hat{\sigma}_{\min}$ for $i = 1, \ldots, h$, which are similar to the adaptive confidence intervals of Section 4.2, but with $\hat{\sigma}_{\min}$ computed from all $h$ estimators $\hat{\beta}_k$ rather than by setting aside $\hat{\beta}_i$ when testing $H_{0,i} : \beta_i = 0$. The more powerful version of this test is discussed in Section 5.2.

## 5.2 *Step-Down Tests*

Although a single-step test compares each effect estimate with the same critical value, a *step-down test* uses this "single-step" critical value only for the largest effect estimate, then "steps down" to test the next largest effect estimate using a sharper critical value, stepping down iteratively and stopping only when an effect is not significant. It is well known, by virtue of sharper critical values after testing the effect with largest estimate, that simultaneous step-down tests have a clear power advantage over simultaneous single-step tests; see, also, Chapter 6.

Although step-up tests are analogous to step-down tests, they are not considered here because error rate control remains an open problem. Step-down or step-up tests have been proposed for the analysis of orthogonal saturated designs by Voss (1988), Voss and Wang (2005), Venter and Steel (1996, 1998), Langsrud and Naes (1998), and Al-Shiha and Yang (1999). Here we provide the details of the tests of Voss and Wang (2005) because they have been proved to control the error rate. The other tests are intuitively attractive but have been "justified" only empirically.

To develop the test, we use the *closed testing procedure* of Marcus et al. (1976). This procedure requires the construction of a size-$\alpha$ test of the hypothesis

$$H_{0,i} : \beta_i = 0, \quad \text{for all } i \in I$$

for each nonempty index set $I \subset \{1, \ldots, h\}$. We test this null hypothesis using the test statistic

$$T_I = \max_{i \in I} T_i, \quad \text{where } T_i = \hat{\beta}_i^2 / \hat{\sigma}_{\min}^2, \tag{7}$$

and where $\hat{\sigma}_{\min}^2$ is defined as in equation (6) but with the modification that each $\hat{\sigma}_j^2$ is computed using all $h$ effect estimators $\hat{\beta}_i$, rather than setting one aside.

Let $c_I$ denote the upper-$\alpha$ quantile of the distribution of $T_I$ when all $h$ effects are zero. Then, the test that rejects $H_{0,I}$ if $T_I > c_I$ is a size-$\alpha$ test of the null hypothesis because the test statistic $T_I$ is a nonincreasing function of $\hat{\beta}_k^2$ for each $K \notin I$ (Voss and Wang, 2005). For each $i$, the closed testing procedure rejects $H_{0,i} : \beta_i = 0$ if and only if $H_{0,I}$ is rejected for each $I$ containing $i$. Use of this procedure controls the simultaneous error rate to be at most $\alpha$; see Marcus et al. (1976). It requires the

testing of $2^h - 1$ hypotheses, one for each subset $I$ of effects. It is then necessary to sort through the results to determine which effects $\beta_i$ can be declared to be nonzero. However, by definition of the test statistic in (7), $I \subset I'$ implies that $T_I \leq T_{I'}$. Also, it can be shown that the critical values $c_I$ decrease as the size of the set $I$ decreases. Thus, we can obtain a shortcut as follows (see also Chapter 6).

**Step-Down Test Procedure:** Let $[1], \ldots, [h]$ be the indices of the effects after reordering so that $T_{[1]} < \cdots < T_{[h]}$. We denote by $c_{j,\alpha}$ the upper-$\alpha$ critical value $c_I$ for any index set $I$ of size $j$. The steps of the procedure are:

S1: If $T_{[h]} > c_{h,\alpha}$, then infer $\beta_{[h]} \neq 0$ and go to step 2; else stop.
S2: If $T_{[h-1]} > c_{h-1,\alpha}$, then infer $\beta_{[h-1]} \neq 0$ and go to step 3; else stop.
S3: $\cdots$

This procedure typically stops within a few steps due to effect sparsity. Voss and Wang (2005) proved, for the above test of $H_{0,i} : \beta_i = 0\,(i = 1, \ldots, h)$, that the probability of making any false inferences (Type I) is at most $\alpha$ for any values of the parameters $\beta_1, \ldots, \beta_h$.

We now apply the step-down test to the plasma etching experiment, choosing a simultaneous significance level of $\alpha = 0.05$ and $J = \{8, 12\}$, as described in Section 4.2. We obtained, via simulation, the values $w_8 = 4.995$, $w_{12} = 2.074$, $c_{15,0.05} = 4.005$, $c_{14,0.05} = 3.969$, and subsequent critical values not needed here. The values of $w_8$ and $w_{12}$ are different from those in Section 4.2 for confidence intervals, because now no $\hat{\beta}_i$ is set aside to obtain $\hat{\sigma}_{\min}$. For testing each effect, $\hat{\sigma}_8^2 = 956.97$ and $\hat{\sigma}_{12}^2 = 1619.94$, so that $\hat{\sigma}_{\min} = 956.97$. Hence, the minimum significant difference for the largest estimate is $c_{15,0.05}\hat{\sigma}_{\min}^2 = 123.89$, and $A$ is declared to be significantly different from zero. Stepping down, the minimum significant difference for the second largest estimate is $c_{14,0.05}\hat{\sigma}_{\min}^2 = 122.78$ and $AB$ is not declared to be significantly different from zero, nor are any of the remaining effects at the simultaneous 5% level of significance.

We recommend the use of this method for the analysis of orthogonal saturated designs. It is the only adaptive step-down test in the literature known to control Type I error rates. Furthermore, it is more powerful than the corresponding single-step test. The single-step test is analogous to the simultaneous confidence intervals of Wang and Voss (2003) and the latter were shown via simulation to provide good minimax and average power over a variety of parameter configurations.

## 6 Discussion

There are important advantages in using adaptive robust procedures that strongly control the error rate. Strong control of the error rate provides the statistical rigor for assessing the believability of any assertions made about the significance of the main effects or interactions, whether confidence intervals or tests are applied. Use of a robust adaptive procedure allows the data to be used efficiently.

From the mathematical viewpoint, the existence of robust adaptive procedures that strongly control the error rates is facilitated by the availability of an adaptive variance estimator that is stochastically smallest when all $\beta_i$ are zero (the null case). Wang and Voss (2003) provided a large class of such robust adaptive estimators, with a lot of flexibility in the possible choices of the set $J$ and weights $w_j$ for construction of adaptive variance estimates. A remaining problem is to determine which estimators in the class are the most robust, in the sense of performing well over all reasonable parameter configurations under effect sparsity. Additional simulation studies seem necessary to investigate this issue; see Wang and Voss (2003).

One way to construct or formulate an adaptive variance estimator is to have multiple variance estimators and to have a data-dependent choice of which one is used. Adaptive methods known to strongly control error rates use special variance estimators of this type. In particular, the construction of adaptive variance estimators that allow strong control of error rates depends fundamentally on choosing between several possible variance estimators where (i) each possible variance estimator is a nonincreasing function of each squared estimator $\hat{\beta}_i^2$, and (ii) the adaptive estimator is the minimum of these contending variance estimators. Under these circumstances, this minimum is also nonincreasing in each squared estimator $\hat{\beta}_i^2$, as required for application of the Stochastic Ordering Lemma. The lemma also requires that the estimators $\hat{\beta}_i$ be independent and that the distribution of each squared estimator $\hat{\beta}_i^2$ be nondecreasing in $\hat{\beta}_i^2$. These requirements all hold for analysis of an orthogonal design under standard linear model assumptions. Under such assumptions, an orthogonal design yields independent effect estimators. A design is *nonorthogonal* if any of the estimators $\hat{\beta}_i$ are correlated under standard linear model assumptions.

For a nonorthogonal design, the problem of how to construct a variance estimator that is robust, adaptive, and known to strongly control error rates is difficult. So far, there is only one procedure known to strongly control error rates. For analysis of nonorthogonal designs, this method, developed by Kinateder et al. (1999), builds upon a variance estimation approach of Kunert (1997) using sequential sums of squares. However, the use of sequential sums of squares is equivalent to making a linear transformation of the estimators $\hat{\beta}_i$ to obtain independent estimators $\hat{\tau}_i$, for which the corresponding sequential sums of squares are also independent. Unfortunately, if the effects are entered into the model in the order $\beta_1, \beta_2, \ldots,$ then the expected value of $\hat{\tau}_i$ can involve not only $\tau_i$ but also the effects $\tau_j$ for $j \geq i$; see Kunert (1997). As a consequence, effect sparsity is diminished for the means of the transformed estimators, $\hat{\tau}_i$.

The problem of data analysis is even harder for supersaturated designs. Then, not only is there necessarily nonorthogonality, but estimability of effects also becomes an issue. Chapter 8 discusses some of the serious problems involved in the analysis of supersaturated designs. Related references include Abraham et al. (1999), Lin (2000), and Holcomb et al. (2003). There is currently no method of analysis of supersaturated designs that is known to provide strong control of error rates. Further work is needed in this area, as well as on the analysis of nonorthogonal saturated designs.

Returning to the discussion of orthogonal saturated designs, Lenth's (1989) method works well in general and no simulation study so far has detected a parameter configuration for which the error rate is not controlled. Also, his procedure is not very dependent on making a good initial guess for the number of negligible effects, although use of the median causes his method to break down if more than half the effects are large. It is of great interest to show that his method strongly controls the error rate. We cannot use the adaptive technique developed by Wang and Voss (2003) to resolve this question for Lenth's method because their method uses a monotone function of the absolute effect estimates to estimate $\sigma_\beta$, whereas Lenth's method and its variants, proposed by Dong (1993) and Haaland and O'Connell (1995), do not.

In the search for nonnegligible effects under effect sparsity, it may seem more reasonable to step up than to step down, that is, to start with evaluation of the smaller estimates, and step up from the bottom until the first significant jump in the estimates is found. Then all effects corresponding to the large estimates could be declared nonzero. Step-down tests, where the largest estimates are considered first and one steps down as long as each estimate in turn is significant, are often justifiable by the closure method of Marcus et al. (1976). However, the mathematical justification of step-up tests, including those of Venter and Steel (1998) and others, remains an interesting and open issue. Wu and Wang (2004) have made some progress in this direction and have provided a step-up test for the number of nonzero effects that provides strong control of error rates. If at least three effects are found to be nonzero, one would like to conclude that the three effects corresponding to the three largest estimates are nonzero, but the procedure does not provide this guarantee. Clearly further research on step-up procedures is needed.

It is appropriate to close this chapter with the following reminder. Although the focus of the work described here is on controlling the Type I error rate for adaptive robust methods, the problem of Type II errors in the analysis of screening experiments should not be overlooked. If a Type II error is made, then an important factor would be ignored in subsequent experiments. On the other hand, if a Type I error is made, then an inactive factor is unnecessarily kept under consideration in future experiments, which is less serious. We argue that the best tests are the most powerful tests of specified size. In a simulation study, we showed (Wang and Voss, 2003) that, in terms of power, adaptive methods known to have strong control of Type I error rates are competitive with alternative methods for which strong control of Type I error rates remains to be established. Hence, one can use adaptive robust methods known to strongly control error rates without sacrificing power.

*Acknowledgments.* The authors are grateful to anonymous reviewers for their helpful comments. Thanks especially to the volume editors. This research was supported by National Science Foundation Grant No. DMS-0308861.

## References


Abraham, B., Chipman, H., and Vijayan, K. (1999). Some risks in the construction and analysis of supersaturated designs. *Technometrics*, **41**, 135–141.

Alam, K. and Rizvi, M. H. (1966). Selection from multivariate normal populations. *Annals of the Institute of Statistical Mathematics*, **18**, 307–318.

Al-Shiha, A. A. and Yang, S. S. (1999). A multistage procedure for analyzing unreplicated factorial experiments. *Biometrical Journal*, **41**, 659–670.

Berk, K. N. and Picard, R. R. (1991). Significance tests for saturated orthogonal arrays. *Journal of Quality Technology*, **23**, 79–89.

Birnbaum, A. (1959). On the analysis of factorial experiments without replication. *Technometrics*, **1**, 343–357.

Box, G. E. P. and Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics*, **28**, 11–18.

Box, G. E. P. and Meyer, R. D. (1993). Finding the active factors in fractionated screening experiments. *Journal of Quality Technology*, **25**, 94–105.

Brown, L. D., Hwang, J. T. G., and Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics*, **26**, 2345–2367.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, **1**, 311–341.

Dean, A. M. and Voss, D. T. (1999). *Design and Analysis of Experiments*. Springer, New York.

Dong, F. (1993). On the identification of active contrasts in unreplicated fractional factorials. *Statistica Sinica*, **3**, 209–217.

Haaland, P. D. and O'Connell, M. A. (1995). Inference for effect-saturated fractional factorials. *Technometrics*, **37**, 82–93.

Hamada, M. and Balakrishnan, N. (1998). Analyzing unreplicated factorial experiments: A review with some new proposals. *Statistica Sinica*, **8**, 1–41.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley and Sons, New York.

Holcomb, D. R., Montgomery, D. C., and Carlyle, W. M. (2003). Analysis of supersaturated designs. *Journal of Quality Technology*, **35**, 13–27.

Kinateder, K. J., Voss, D. T., and Wang, W. (1999). Exact confidence intervals in the analysis of nonorthogonal saturated designs. *American Journal of Mathematical and Management Sciences*, **20**, 71–84.

Kunert, J. (1997). On the use of the factor-sparsity assumption to get an estimate of the variance in saturated designs. *Technometrics*, **39**, 81–90.

Langsrud, O. and Naes, T. (1998). A unified framework for significance testing in fractional factorials. *Computational Statistics and Data Analysis*, **28**, 413–431.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, **31**, 469–473.

Lin, D. K. J. (2000). Recent developments in supersaturated designs. In *Statistical Process Monitoring and Optimization*. Editors: S. H. Park and G. G. Vining, pages 305–319. Marcel Dekker, New York.

Mahamunulu, D. M. (1967). Some fixed-sample ranking and selection problems. *Annals of Mathematical Statistics*, **38**, 1079–1091.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

Venter, J. H. and Steel, S. J. (1996). A hypothesis-testing approach toward identifying active contrasts. *Technometrics*, **38**, 161–169.

Venter, J. H. and Steel, S. J. (1998). Identifying active contrasts by stepwise testing. *Technometrics*, **40**, 304–313.

Voss, D. T. (1988). Generalized modulus-ratio tests for analysis of factorial designs with zero degrees of freedom for error. *Communications in Statistics: Theory and Methods*, **17**, 3345–3359.

Voss, D. T. (1999). Analysis of orthogonal saturated designs. *Journal of Statistical Planning and Inference*, **78**, 111–130.

Voss, D. T. and Wang, W. (1999). Simultaneous confidence intervals in the analysis of orthogonal saturated designs. *Journal of Statistical Planning and Inference* **81**, 383–392.

Voss, D. T. and Wang, W. (2005). On adaptive testing in orthogonal saturated designs. *Statistica Sinica*, in press.

Wang, W. and Voss, D. T. (2001a). Control of error rates in adaptive analysis of orthogonal saturated designs. *Annals of Statistics*, **29**, 1058–1065.

Wang, W. and Voss, D. T. (2001b). On the analysis of nonorthogonal saturated designs using effect sparsity. *Statistics and Applications* **3**, 177–192.

Wang, W. and Voss, D. T. (2003). On adaptive estimation in orthogonal saturated designs. *Statistica Sinica*, **13**, 727–737.

Wu, S. S. and Wang, W. (2004). Step-up simultaneous tests for identifying active effects in orthogonal saturated designs. Technical report 2004-018, Department of Statistics, University of Florida.

Ye, K. Q. and Hamada, M. (2000). Critical values of the Lenth method for unreplicated factorial designs. *Journal of Quality Technology*, **32**, 57–66.

Zahn, D. A. (1969). An empirical study of the half-normal plot. PhD thesis, Harvard University, Boston.

Zahn, D. A. (1975a). Modifications of and revised critical values for the half-normal plots. *Technometrics*, **17**, 189–200.

Zahn, D. A. (1975b). An empirical study of the half-normal plot. *Technometrics*, **17**, 201–211.

# 13
# Screening for the Important Factors in Large Discrete-Event Simulation Models: Sequential Bifurcation and Its Applications

Jack P. C. Kleijnen, Bert Bettonvil and Fredrik Persson

Screening in simulation experiments to find the most important factors, from a very large number of factors, is discussed. The method of sequential bifurcation in the presence of random noise is described and is demonstrated through a case study from the mobile telecommunications industry. The case study involves 92 factors and three related, discrete-event simulation models. These models represent three supply chain configurations of varying complexity that were studied for an Ericsson factory in Sweden. Five replicates of observations from 21 combinations of factor levels (or scenarios) are simulated under a particular noise distribution, and a shortlist of the 11 most important factors is identified for the most complex of the three models. Various different assumptions underlying the sequential bifurcation technique are discussed, including the role of first- and second-order polynomial regression models to describe the response, and knowledge of the directions and relative sizes of the factor main effects.

## 1 Introduction

In this chapter, we explain the technique of sequential bifurcation and add some new results for random (as opposed to deterministic) simulations. In a detailed case study, we apply the resulting method to a simulation model developed for Ericsson in Sweden. In Sections 1.1 to 1.3, we give our definition of screening, discuss our view of simulation versus real-world experiments, and give a brief indication of various screening procedures.

Our case study is introduced in Section 1.4. The assumptions behind, and the steps involved in, sequential bifurcation are described in Sections 2.2 and 2.3 and the steps are illustrated using a first-order polynomial model with random noise. In Section 2.4, a more realistic model involving interactions is used for screening the important factors in the case study. Issues of programming are addressed in Section 3.

### *1.1 A Definition of Screening*

We define *screening* as the "search for the most important factors among a large set of factors in an experiment." For example, in our case study described in

Section 1.4, 92 factors are studied but we find only 11 of these to be important. This is in line with the principle of effect sparsity, see Chapters 1 and 8. The simplest definition of "importance" occurs when an experiment has a single response (output from computer code) and the factors have only additive effects; that is, the input–output relation is modelled by a first-order polynomial in regression terminology or a "main effects only" model in analysis of variance terminology (also see Chapter 8). The most important factor is then the one that has the largest absolute value for its first-order effect or main effect; the least important factor is the one whose effect is closest to zero.

The goal of screening is to draw up a shortlist of important factors from a long list of potentially important factors. Depending on the application, this shortlist might lead to a more thorough investigation of the possibly important factors via additional experiments (Kleijnen et al., 2002) or through an optimization and robustness analysis (Kleijnen et al., 2003). In an ecological case study, Bettonvil and Kleijnen (1996) identified a shortlist of factors which included some factors that the ecological experts had not expected to have large effects!

It is also important to find out from the screening experiment which factors are "certainly" unimportant so that the clients of the simulation analysts are not bothered by details about these factors. We are distinguishing, in this chapter, between the simulation analysts, who develop a simulation model and run experiments on this model (as, for example, in Chapter 14) and their clients, who are the managers and other users of the real system being simulated.

Of course, the perceived importance of factors depends on the *experimental domain* or *design region* which is the experimental area to be explored and is also called the "experimental frame" by Zeigler et al. (2000). The clients must supply information about this domain to the simulation analysts, including realistic ranges of the individual factors and limits on the admissible scenarios or combinations of factor levels; for example, in some applications the factor values must add up to 100%.

We view the real or the simulated system as a black box that transforms inputs into outputs. Experiments with such a system are often analyzed through an approximating regression or analysis of variance model. Other types of approximating models include those for Kriging, neural nets, radial basis functions, and various types of splines. We call such approximating models *metamodels*; other names include auxiliary models, emulators, and response surfaces. The simulation itself is a model of some real-world system. The goal is to build a parsimonious metamodel that describes the input–output relationship in simple terms.

We emphasize the following chicken-and-egg problem: once the design for the simulation experiment is specified and the observations have been obtained, parameters for an appropriate metamodel can be estimated. However, the types of metamodels that the analyst desires to investigate should guide the selection of an appropriate design.

## *1.2 Simulation Versus Real-World Experiments*

The classic design of experiments focuses on real-world experiments; see, for example, the classic textbook by Box et al. (1978) or the recent textbook by

Myers and Montgomery (2002). We, however, focus on experiments with computer or simulation models which may be either deterministic or stochastic; also see Kleijnen et al. (2002). An introduction to simulation modelling is provided by Law and Kelton (2000).

In simulation experiments, using advances in computing power, analysts are no longer bound by some of the constraints that characterize real-world experiments. This is a challenge, as it requires a new mindset. We argue that the ways in which simulation experiments and real-world experiments should be approached are fundamentally different, especially in the following three aspects.

(i) In real-world experiments, the analysts must often select a design that is executed in one shot (for example, one growing season in an agricultural experiment). In simulation experiments, however, the data are collected sequentially because a standard computer operates sequentially and the use of computers in parallel is still an exception. Thus the analysts may start with a small design for a very simple metamodel, then test (validate) the adequacy of that model and, only if that model is rejected, need they augment the original design to enable the estimation of a more complicated model. This is a two-stage design. In this chapter, however, we present an alternative strategy in which the design is analyzed for each new observation before the next design point is selected (see also Kleijnen et al., 2002; Kleijnen and Sargent, 2000).

On the occasions when analysts must collect observations sequentially in real-world experiments, the experiment is viewed as prone to validity problems. Hence, the analysts randomize the order in which the factor level combinations are observed to guard against time-related changes in the experimental environment (such as temperature, humidity, consumer confidence, and learning effects) and perform appropriate statistical tests to determine whether the results have been contaminated. For the simulation experiment, on the other hand, an input file can be generated once a particular design type has been chosen. Such a file can be executed sequentially and efficiently in batch mode without human intervention and the computer implements the sequential design and executes rules for selecting the next design point based on all preceding observations.

(ii) In real-world experiments, typically only a few factors are varied. In fact, many published experiments deal with fewer than five factors. The control of more than, say, ten factors in real-world experiments is a challenging area; see, for example, Chapters 4, 8, and 9. In simulation experiments, the computer code typically involves a very large number of factors; for example, there are 92 factors in our case study. Good computer programming avoids the need to fix the values of any of these factors within the code and allows them to be read in from an input file. Thus, other than checking that the factor level combinations give a scenario within the experimental domain, there are no restrictions on the scenarios that can be run. Such a practice can automatically provide a long list of potential factors. Analysts should confirm whether they, indeed, wish to experiment with all of these factors or whether they wish a priori to fix some factors at nominal (or base) levels. This type of coding helps to unfreeze the mindset of users who might otherwise be inclined to focus on only a few factors to be varied in the experiment. For example, Persson and Olhager (2002) simulated only nine combinations of factor

levels. An investigation of a large set of factors, however, implies that computer time may become an issue and then special screening procedures, such as sequential bifurcation, and optimized computer code, must be used. For example, in our case study, one run of the simulation code originally took three hours; this was reduced to 40 minutes after modification of our code.

(iii) *Randomness or random variability* occurs in real-world experiments because the experimenters cannot control all of the factors that affect the response; for example, human participants in an experiment have ambitions that cannot be fully controlled by the experimenters. In computer simulation, such effects can be modelled through random input variables; for example, the arrivals of customers may be modelled as a Poisson process so that the times between two successive arrivals are exponentially distributed. Values for random variables are generated through pseudorandom numbers. A single simulation run gives an observation on each of the responses of interest to the analysts; for example, in our supply chain case study of Section 1.4, one simulation run represents the operations of a supply chain during 17 weeks and the main response is the average cost per week. To obtain independently and identically distributed observations, the analysts generate several simulation runs or *replicates*, which all start in the same initial state of the simulated system but use different pseudorandom numbers to generate the values of the random input variables.

## *1.3 Screening Procedures for Simulation*

Campolongo et al. (2000) discussed a variety of screening methods for simulation, as opposed to real-world experiments. These include one-factor-at-a-time, such as the method of Morris (1991), iterated fractional factorial designs of Andres and Hajas (1993) as well as sequential bifurcation. The authors refer to available software and emphasize key assumptions. Some of the methods, including sequential bifurcation (explained in Section 2), require fewer observations than there are factors, as in supersaturated designs used for single shot experiments (see Chapter 8; also Westfall et al., 1998). Recently, De Vos et al. (2003) applied multi-stage group-screening to a model using the @Risk software. This software is distinct from the discrete-event dynamic simulation software, such as the Taylor II software (see Incontrol, 2003) used in our case study. Various other screening methods are presented in this book; in particular in the closely related Chapters 9 and 14.

## *1.4 The Ericsson Case Study: Supply Chain Simulation*

We now discuss a recent example of simulation that models three alternative configurations for a supply chain in the mobile communications industry at the Ericsson company in Sweden. A *supply chain* consists of several *links* which may be separate companies or independent business units of a single large company. Examples of links are retailers, wholesalers, distributors, and factories. Customary strategies imply that, at the individual links of the supply chain, decisions are made based on

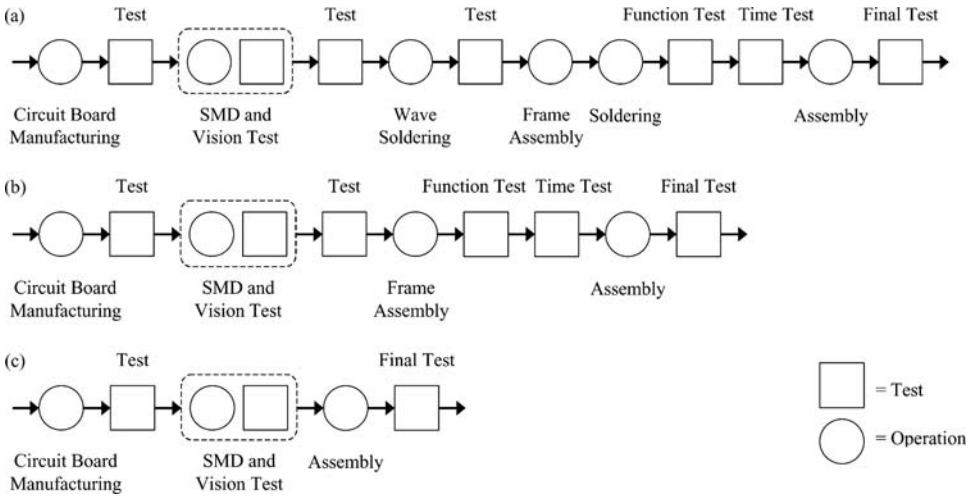



FIGURE 1. The three supply chain configurations: (a) the Old, (b) the Current, (c) the Next Generation.

information about their *nearest neighbour* links instead of information about the *final* link. The final link deals with the ultimate customers who demand the final product, rather than some intermediate product or raw material. Under these customary strategies, if the final demand at the final link increases by, say, 1%, then there is likely to be a much greater increase in the orders placed from a link to its predecessor when the link is early in the chain; this is known as the *bullwhip effect*. Banks et al. (2002) and Kleijnen (2005) discussed simulation in supply chain management—a rapidly developing field in operations research; see, for example, the *Proceedings of the 2004 Winter Simulation Conference* and Simchi-Levi et al. (2000).

A central issue in supply chain management is the *lean and mean* strategy which is the improvement in the performance of the total chain through the elimination of links, or steps within links. In the Ericsson case study, three supply chain configurations are investigated. Each configuration is actually the same chain, but simplified over time. The Old configuration, which existed in 1998, consists of many operations and test stations. The Current (1999) configuration has fewer operational steps. The Next Generation chain is a future configuration that has a minimum of operations and tests.

Figure 1 shows diagrams of the three configurations. A square denotes a test of the products produced by the preceding operation, which is denoted by a circle. There are several types of tests and operations, as the names in the figure show. Products flow through the chain from left to right in the figure. The chain starts with the purchase of "raw" products; next, these products are processed; the chain finishes with the assembly of components into final products, which are then sold. The abbreviation SMD stands for "surface mounted devices", indicating electronic devices that are mounted on the surface of a circuit board (which is

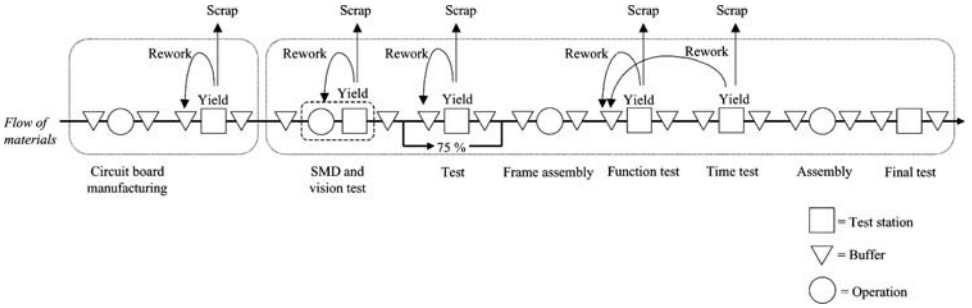



FIGURE 2. The simulation model of the Current supply chain configuration.

modern technology compared with devices mounted in holes on the board). In Figure 1, the dotted boxes indicate that the mounting and testing of the circuit board is performed by the same machine—a very fast mounting machine with integrated vision control of device placements. Into this machine is also integrated the heating to make the soldering paste melt and connect the board to the electronic device.

Figure 2 shows the Current simulation model. This figure illustrates that buffers (inventories) are located before and after each test station and operation; products are transported between all operations and test stations.

As described in Section 1.1, the goal of the simulation study is to quantify the relationships between the simulation outputs and the inputs or factors. For this case study, the outputs are the steady-state mean costs of the whole supply chain (discussed in Section 3.2) and the inputs are factors such as lead-time, quality, operation time of an individual process, and number of resources. Our ultimate goal (as reported by Kleijnen et al., 2003) is to find *robust* solutions for the supply chain problem. Thus we distinguish between two types of factors:

1. Factors that are controllable by the company; for example, Ericsson can manipulate the manufacturing processes and select logistic partners for transportation.
2. Noise factors that are uncontrollable and determined by the environment; examples include demand for products, process yield or percentage of faulty products, and scrap percentage at each test station.

The simulation model of the Old supply chain has 92 factors, whereas those of the Current and Next Generation supply chains have 78 and 49 factors, respectively. Details on the results of the simulation are given in Section 3.

## 2 Sequential Bifurcation

Originally, sequential bifurcation was developed in the doctoral dissertation of Bettonvil (1990) which was later summarized by Bettonvil and Kleijnen (1996) and updated by Campolongo et al. (2000) to include a discussion of applications. Other authors have also studied sequential bifurcation (see, for example, Cheng,

1997; Cheng and Holland, 1999). Sequential bifurcation is related to binary search (Wan et al., 2004) which searches a sorted array by repeatedly dividing the search interval in half, beginning with an interval covering the whole array. First, we give an outline of the sequential bifurcation procedure (Section 2.1). Second, we present the assumptions and notation of sequential bifurcation (Section 2.2), and, third, we illustrate the procedure through our case study (Section 2.3).

## 2.1 An Outline of the Sequential Bifurcation Procedure

A formal description of sequential bifurcation can be found in Bettonvil (1990). The procedure follows a sequence of steps. It begins by placing all factors into a single group and testing whether the group of factors has an important effect. If it does, the group is split into two subgroups and each of these is then tested for importance. The procedure continues in this way, discarding unimportant groups and splitting important groups into smaller groups. Eventually, all factors that are not in groups labelled as unimportant are tested individually.

## 2.2 Assumptions and Notation of Sequential Bifurcation

The two basic assumptions of sequential bifurcation are as follows.

*Assumption 1*: an adequate metamodel is a first-order polynomial, possibly augmented with two-factor interactions; that is,

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{1:2} x_1 x_2 + \cdots + \beta_{(k-1):k} x_{k-1} x_k + \epsilon, \qquad (1)$$

where the following notation is used:

$Y$: The response for the metamodel
$k$: The total number of factors in the experiment
$\beta_j$: The first-order or main effect of factor $j$ with $j = 1, \ldots, k$
$\beta_{j':j}$: The interaction effect of the factors $j'$ and $j$ with $1 \leq j' < j \leq k$
$x_j$: The value of the $j$th factor, standardised to lie in $[-1, +1]$
$\epsilon$: The noise variable arising from both the use of pseudorandom numbers and approximation error

We observe that this metamodel is linear in the parameters $\beta_j$ and $\beta_{j':j}$ but nonlinear in the variables $x_j$. At the end of the case study in Section 3, we try to validate the assumption that the model is adequate.

To estimate the parameters in the simple metamodel (1), it is most efficient to experiment with *only two* levels (values) per factor. In practice, it is important that these levels are chosen to be realistic, so the users of the underlying simulation model should provide these values.

*Assumption 2*: the signs of all main effects are known and are nonnegative, that is,

$$\beta_j \geq 0 \quad (j = 1, \ldots, k). \qquad (2)$$

We need this assumption because, otherwise, main effects might cancel each other (see also Chapter 9). Our experience is that, in practice, this assumption is easy to satisfy; that is, it is straightforward to define the upper and lower level of each factor such that changing a factor from its lower to its upper level does not decrease the expected response. For example, in our case study, some factors refer to transportation speed: the higher this speed, the lower the work in process (WIP) and hence the lower the costs. Other examples are provided by Lewis and Dean (2001). Consequently, we call the two levels of each factor the *low level* and the *high level*, respectively, where the low level results in the lower expected (simulation) response and the high level produces the higher expected response. Nevertheless, if, in a particular case study, Assumption 2 seems hard to meet for specific factors, then these factors should be treated "individually"; that is, none of these factors should be grouped with other factors in the sequential bifurcation procedure.

The simplest experimental domain is a $k$-dimensional hypercube where the $j$th original factor is coded such that its respective low and high values, $l_j$ and $h_j$, correspond to the values $-1$ and $+1$ of the corresponding standardised factor. Thus, the level $z_j$ of the original factor is transformed to level $x_j$ of the standardised factor, where

$$x_j = \frac{z_j - (h_j + l_j)/2}{(h_j - l_j)/2}. \tag{3}$$

The scaling in (3) ensures that the experimental results are insensitive to the scale used. Therefore, we may rank or sort the factors by the size of their main effects so that the most important factor is the one with the highest main effect, and so on. We note that the larger the range of an untransformed factor, the larger is the difference between the responses at the two levels and the larger is the main effect of the transformed factor. (See, also, the "unit cost" effects of Cheng and Holland, 1999.)

For our case study, we could not obtain information on the factor ranges from Ericsson. We decided therefore to change most factors by 5% of the base values reported for the existing system by Persson and Olhager (2002) and to change the transportation speeds between operations by 25% (see Cheng and Holland, 1999).

The *efficiency* of sequential bifurcation, as measured by the number of observations (that is, simulation runs and hence simulation time), increases if the individual factors are renumbered to be in increasing order of importance (see Bettonvil 1990, page 44), so that

$$\beta_{j'} \leq \beta_j \quad (j' < j). \tag{4}$$

We try to realize this efficiency gain by applying prior knowledge about the factors in the simulated real system. In the case study, we anticipated that the environmental factors would be the most important, so these factors appear last in the list of factors.

In order to increase the efficiency further, it is helpful to use any available knowledge about the simulated real system to keep *similar* factors together. For example, we group together all "test yield" factors and conjecture that, if one

yield factor is unimportant, then all yield factors are likely to be unimportant too. Bettonvil (1990, pages 40–43) further divided factor groups so that the number of factors per resulting subgroup is a power of two. We use his approach as a secondary guideline, unless it implies splitting up a group of related factors. (Cheng, 1997, splits groups into two subgroups of equal size.) In our sequential bifurcation for the Old supply chain, for example, we split the first 49 factors into a group of 32 ($=2^5$) factors and a group of the remaining factors. Figure 3 shows the results of applying sequential bifurcation to our Ericsson case study, where the factors are labeled 1–92, and the continuous horizontal band at the top of the figure indicates that all factors have been grouped together in the first step. The steps of the sequential bifurcation procedure are given in Section 2.3. We assume initially that a first-order polynomial is an adequate metamodel, so that the interaction parameters in (1) are zero, and also assume that the expected value of the noise variable $\epsilon$ is zero; that is,

$$\beta_{j':j} = 0 \quad (j' \neq j) \quad \text{and} \quad \mu_\epsilon = 0. \tag{5}$$

We introduce the following additional sequential bifurcation notation adapted for replicated random responses. We use $y_{(j);r}$ to represent the observed (simulation) output in replicate $r$ when factors 1 to $j$ are set at their high levels and the remaining factors are set at their low levels ($r = 1, \ldots, m$).

We define $\beta_{j'-j}$ to be the sum of the main effects for factors $j'$ to $j$; that is,

$$\beta_{j'-j} = \sum_{i=j'}^{j} \beta_i. \tag{6}$$

An estimate of this *aggregated main effect* $\beta_{j'-j}$, using only the output from replicate $r$, is

$$\hat{\beta}_{j'-j;r} = \frac{y_{(j);r} - y_{(j'-1);r}}{2}. \tag{7}$$

The sequential bifurcation procedure starts by observing (simulating) the two most extreme scenarios. In scenario 1, all factors are at their low levels and, in scenario 2, all factors are at their high levels. From the metamodel (1), we obtain the expected values of the response variables as

$$E(Y_{(0)}) = \beta_0 - \beta_1 - \cdots - \beta_k,$$
$$E(Y_{(k)}) = \beta_0 + \beta_1 + \cdots + \beta_k.$$

It follows that

$$E(Y_{(k)}) - E(Y_{(0)}) = 2(\beta_1 + \cdots + \beta_k) \tag{8}$$

which shows that the estimator based on (7) is unbiased.

For the individual main effect of the $j$th factor, the estimate from the $r$th replicate is

$$\hat{\beta}_{j;r} = \frac{y_{(j);r} - y_{(j-1);r}}{2}. \tag{9}$$

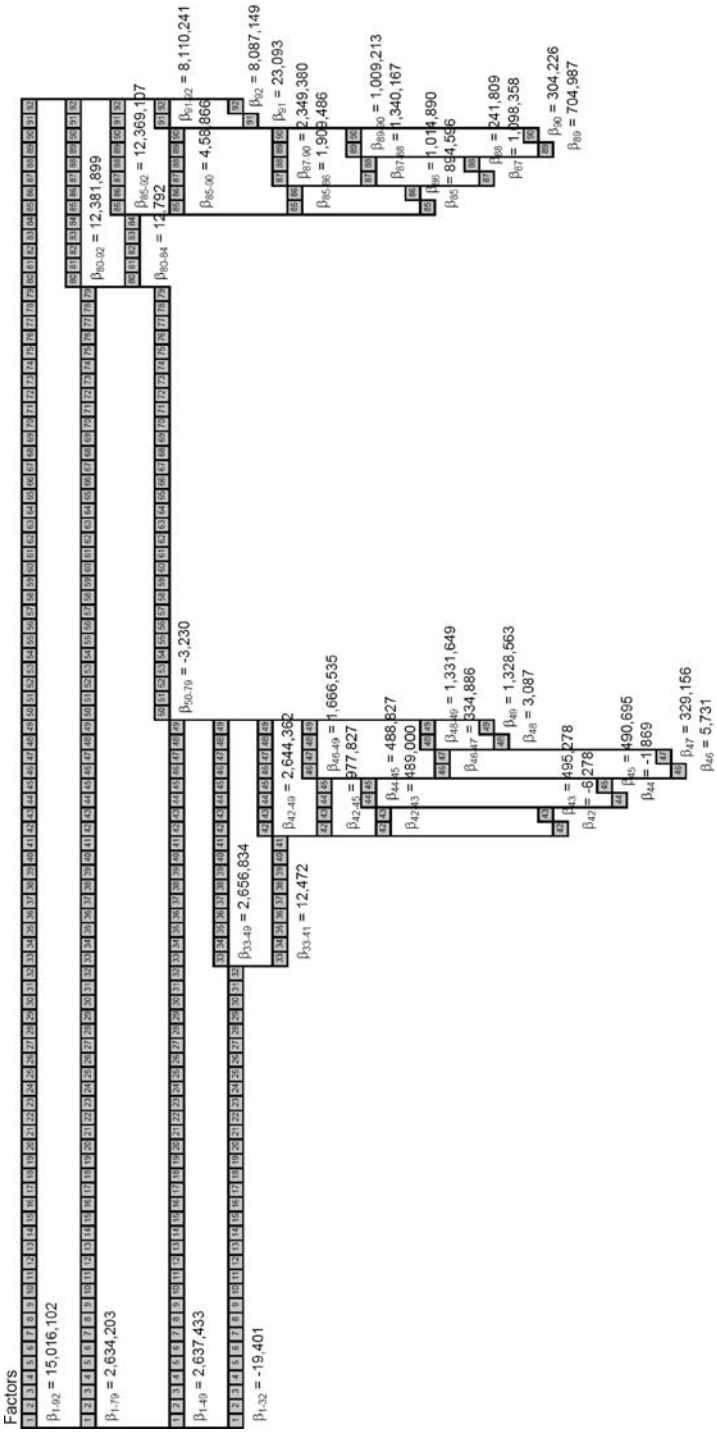



FIGURE 3. The steps of sequential bifurcation applied to the Old supply chain configuration, assuming a first-order polynomial metamodel. Estimates of the $\beta_{j'-j}$ defined in equation (6) are indicated at each step.

From the $m$ replicates, we compute the sample average and its sample variance for each (aggregated or individual) estimated main effect. For example, for the individual main effect of factor $j$, we obtain

$$\bar{\hat{\beta}}_j = \frac{\sum_{r=1}^{m} \hat{\beta}_{j;r}}{m} \quad \text{and} \quad s^2(\bar{\hat{\beta}}_j) = \frac{\sum_{r=1}^{m} (\hat{\beta}_{j;r} - \bar{\hat{\beta}}_j)^2}{m(m-1)}. \tag{10}$$

The variance estimators $s^2(\bar{\hat{\beta}}_j)$ of the estimated effects $\bar{\hat{\beta}}_j$ allow unequal response variances and the use of *common* pseudorandom numbers. This is a well-known technique used in simulation experiments to reduce the variances of the estimated factor effects (Law and Kelton, 2000). This technique uses the same pseudorandom numbers when simulating the system for different factor combinations, thus creating positive correlations between the responses. Consequently, the variances of the estimated effects are reduced. This technique is similar to blocking in real-world experiments; see, for example, Dean and Voss (1999, Chapter 10).

## 2.3 The Steps of Sequential Bifurcation

We now illustrate the sequential bifurcation procedure using the Old simulation model which has $k = 92$ factors and $m = 5$ replicates. Table 1 gives the observations for the two extreme scenarios for each replicate. We start sequential bifurcation by finding the average simulated response when all factors are at their low levels, $\bar{y}_{(0)} = 3{,}981{,}627$, and that when all factors are at their high levels, $\bar{y}_{(92)} = 34{,}013{,}832$, where the overline denotes the average computed from the $m = 5$ replicates. So, the estimated effect of all 92 factors aggregated together is obtained from (7), together with a formula analogous to (10), as $\bar{\hat{\beta}}_{1-92} = (34{,}013{,}832 - 3{,}983{,}627)/2 = 15{,}016{,}102$. This estimate is shown in Figure 3 immediately below the first shaded line listing all factor labels from 1 through 92. The standard error of this estimated aggregated effect, averaged over the replicates, is $s(\bar{\hat{\beta}}_{1-92}) = 94{,}029.3/\sqrt{5} = 42{,}051.18$.

To *test* the importance of the estimated (either aggregated or individual) main effects statistically, we assume that the (simulated) outputs for each scenario are approximately normally and independently distributed. Different scenarios may produce observations with different variances; the use of common pseudorandom

TABLE 1. Observations for the first two scenarios simulated in sequential bifurcation for the Old supply chain

| Replicate | $y_{(0)}$ | $y_{(92)}$ | $\hat{\beta}_{1-92}$ |
|---|---|---|---|
| 1 | 3,954,024 | 34,206,800 | 15,126,388.0 |
| 2 | 3,975,052 | 33,874,390 | 14,949,669.0 |
| 3 | 3,991,679 | 33,775,326 | 14,891,823.5 |
| 4 | 4,003,475 | 34,101,251 | 15,048,888.0 |
| 5 | 3,983,905 | 34,111,392 | 15,063,743.5 |
| Average | 3,981,627 | 34,013,832 | 15,016,102.4 |
| Standard Error | 18,633 | 180,780 | 94,029.3 |

numbers for different scenarios may produce correlated observations. However, in view of the fact that only large effects need to be detected, we apply a two-sample $t$-test as an approximate test, ignoring variance heterogeneity when determining the degrees of freedom (see Kleijnen, 1987, pages 14–23). We apply a one-sided test because we assume that all individual main effects are nonnegative and rank ordered as in (2) and (4). Using a one-sided two-sample $t$-test with level $\alpha = 0.05$ and 8 degrees of freedom, we conclude that the sum of the 92 main effects is significantly different from zero. Our heuristic uses a fixed $t$-value throughout the whole sequential bifurcation procedure with no adjustment for multiple testing; see Kleijnen (1987, pages 41–45). In practice, significance is not essential but importance is—we are searching for a shortlist of important factors. In a recent paper, Wan et al. (2004) discuss the use of multiple testing procedures in a sequential bifurcation setting.

In hindsight, we might have used fewer replications in the early steps of the procedure, as these steps have higher signal/noise ratio due to the fact that the signal decreases as a result of less aggregation of main effects as the sequential bifurcation progresses.

The aggregated or group effect is an *upper limit U* for the value of any individual main effect. The goal of sequential bifurcation is to find the most "important" factors, that is, the factors that have significant main effects. If, however, we terminate our screening prematurely (for example, because the computer breaks down or our clients get impatient), then sequential bifurcation still allows identification of the factors with the largest main effects.

The *next step* is to divide the current group of 92 factors into *two* subgroups; this explains the term *bifurcation*. Into one subgroup we place all the 79 controllable factors and into the other subgroup we put all 13 environmental factors, as indicated by the second shaded line in Figure 3. Simulation of the five replicates of this scenario gives the average response $\bar{y}_{(79)} = 9{,}250{,}034$ with standard error 14,127. This value, $\bar{y}_{(79)}$, lies between $\bar{y}_{(0)}$ and $\bar{y}_{(92)}$, in line with the sequential bifurcation assumptions. Comparison of $\bar{y}_{(79)}$ and $\bar{y}_{(0)}$ via (9) and (10) gives $\hat{\bar{\beta}}_{1-79} = 2{,}634{,}203$ (with standard error 16,534). Similarly, a comparison of $\bar{y}_{(92)}$ and $\bar{y}_{(79)}$ gives $\hat{\bar{\beta}}_{80-92} = 12{,}381{,}899$ (with standard error 128,220). So, this step splits the total effect $\hat{\bar{\beta}}_{1-92} = 15{,}016{,}102$ of the first step into its two additive components. This step decreases the upper limit $U$ for any individual effect in the first subgroup to 2,634,203; for any individual effect in the second subgroup this limit is 12,381,899.

To decide *where* to split a group into two subgroups, we use several principles which are intended to increase the efficiency of our procedure, as explained in the discussion between (4) and (5). Figure 3 also shows our remaining sequential bifurcation steps. We do not split a group any further when its estimated aggregated main effect is either nonsignificant or negative. For example, the estimated aggregated main effect of factors 50 through 79 is $-3230$ with standard error 31,418.

For this case study, sequential bifurcation stops after 21 steps. The upper limit, denoted by $U(21)$, for the main effect of any remaining individual factor is then

TABLE 2. Important factors identified in sequential bifurcation for the Old supply chain: N/A denotes a dummy factor and $\surd$ denotes an important factor

| Factor | | Model | | |
|---|---|---|---|---|
| | | Old | Current | Next generation |
| 92 | Demand | $\surd$ | $\surd$ | $\surd$ |
| 90 | Yield (circuit board) | $\surd$ | $\surd$ | $\surd$ |
| 89 | Yield (vision test) | $\surd$ | $\surd$ | $\surd$ |
| 88 | Yield (SMD test) | $\surd$ | $\surd$ | N/A |
| 87 | Yield (test, after wave soldering) | $\surd$ | N/A | N/A |
| 86 | Yield (function test) | $\surd$ | $\surd$ | $\surd$ |
| 85 | Yield (time test) | $\surd$ | $\surd$ | N/A |
| 49 | Transportation (between factories) | $\surd$ | $\surd$ | $\surd$ |
| 47 | Transportation (internal, circuit board factory) | $\surd$ | $\surd$ | $\surd$ |
| 45 | Transportation (between SMD and test) | $\surd$ | $\surd$ | $\surd$ |
| 43 | Transportation (between wave soldering and test) | $\surd$ | N/A | N/A |

reduced to 12,792. Our list of the 11 most important factors is given in Table 2. The corresponding reordered list in the left-hand column of Table 3 shows that factor 92 is the most important with an estimated main effect of 8,087,149, and factor 88 has the smallest main effect in the shortlist with estimate 241,809. Remembering that we tried to label the factors such that equation (4) is satisfied, we now conclude that, indeed, factor 92 is the most important factor and that no factor labelled with a number smaller than 43 is significant.

In the next section we consider a more realistic metamodel that includes interactions and we illustrate the design and analysis of experiments for sequential bifurcation under such models.

TABLE 3. The important factors found from sequential bifurcation under two meta-models for the Old supply chain simulation. Details of the factors are given in Table 2

| First-order polynomial model | | | |
|---|---|---|---|
| Without interactions | | With interactions | |
| Factor | Estimated main effect | Factor | Estimated main effect |
| 92 | 8,087,149 | 92 | 4,967,999 |
| 49 | 1,328,563 | 49 | 1,953,819 |
| 45 | 490,695 | 45 | 1,586,534 |
| 43 | 495,278 | 43 | 1,581,504 |
| 87 | 1,098,358 | 87 | 1,057,414 |
| 85 | 894,596 | 85 | 1,054,181 |
| 86 | 1,014,890 | 86 | 1,053,766 |
| 89 | 704,987 | 89 | 599,892 |
| 47 | 329,156 | 47 | 480,914 |
| 90 | 304,226 | 90 | 271,414 |
| 88 | 241,809 | 88 | 181,517 |

## 2.4 Two-Factor Interactions and Foldover Designs in Sequential Bifurcation

Aggregated main effects can be estimated independently of two-factor interactions if the *foldover* principle (Box and Wilson, 1951) is used with sequential bifurcation. Foldover designs are discussed in Chapters 1 and 9 and Kleijnen (1987, page 303). Such designs consist of a set of factor level combinations and their mirror images, that is, with levels switched from high to low, and vice versa. We let $y_{-(j);r}$ denote the observed output with the factors 1 through $j$ set to their low levels in replication $r$ and the remaining factors set to their high levels so that $y_{-(j);r}$ is the mirror observation of $y_{(j);r}$. As an example, we consider the third shaded line in Figure 4. The observations $y_{(49);r}$ and $y_{(-49);r}(r = 1, \ldots, 5)$ are simulated and, from the metamodel (1), it follows that

$$E(Y_{-(49)}) = \beta_0 - \beta_1 - \cdots - \beta_{49} + \beta_{50} + \cdots + \beta_{92} + \beta_{1;2} + \cdots + \\ + \beta_{48;49} - \beta_{1;50} - \cdots - \beta_{49;92} + \beta_{50;51} + \cdots + \beta_{91;92} \quad (11)$$

and

$$E(Y_{(49)}) = \beta_0 + \beta_1 + \cdots + \beta_{49} - \beta_{50} - \cdots - \beta_{92} + \beta_{1;2} + \cdots + \\ + \beta_{48;49} - \beta_{1;50} - \cdots - \beta_{49;92} + \beta_{50;51} + \cdots + \beta_{91;92}. \quad (12)$$

Subtracting these two equations demonstrates that all interactions cancel out. In a similar way, the group effect estimates

$$\hat{\beta}_{j'-j;r} = \frac{(y_{(j);r} - y_{-(j);r}) - (y_{(j'-1);r} - y_{-(j'-1);r})}{4} \quad (13)$$

are unbiased by two-factor interactions, as are the individual main effect estimates

$$\hat{\beta}_{j;r} = \frac{(y_{(j);r} - y_{-(j);r}) - (y_{(j-1);r} - y_{-(j-1);r})}{4} \quad (14)$$

(see Bettonvil, 1990, for more details).

Sequential bifurcation may give misleading results if, say, two factors have unimportant main effects but the interaction between them is important (see, also, Lewis and Dean, 2001). However, we only consider situations in which the following "strong heredity" assumption of Wu and Hamada (2000) holds.

*Assumption 3*: if a factor has no important main effect, then this factor does not interact with any other factor:

$$\beta_j = 0 \Longrightarrow \beta_{j;j'} = 0 \quad (j \neq j'). \quad (15)$$

If, a priori, we suspect that this assumption is violated, then we should investigate such a factor after the screening phase.

The foldover design does *not* enable us to estimate *individual* interactions, but it does enable us to estimate whether interactions are important, as follows. We estimate the main effects from the original scenarios as in Section 2.3, ignoring the mirror scenarios. If the analyses of the foldover design and of the "original"

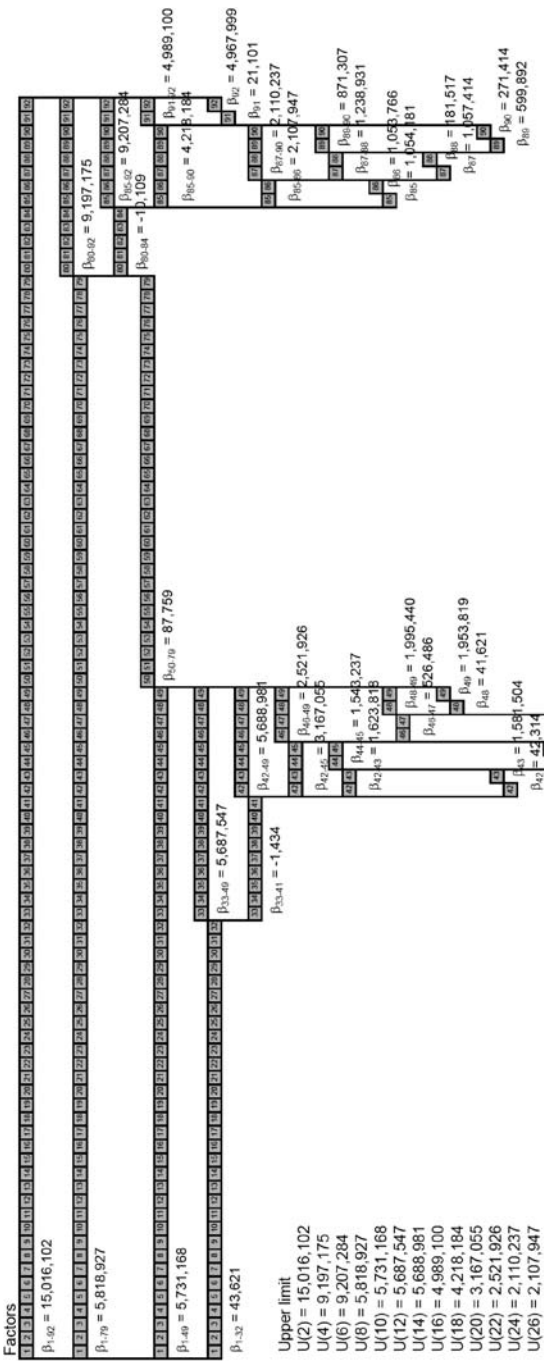



FIGURE 4. Sequential bifurcation assuming a first-order polynomial plus two-factor interactions metamodel, applied to the Old supply chain simulation; includes upper limits for parameter values.

design give the same conclusions, then it can be concluded that interactions are unimportant. This happened, for example, in the ecological simulation reported by Bettonvil (1990) and Bettonvil and Kleijnen (1996). In the present case study, however, we show that interactions are important. (In a follow-up experiment that included only the factors declared to be important, Kleijnen et al., 2003, estimated the sizes of the individual interactions from a Resolution V design.)

Analogous to Figure 3, Figure 4 shows the sequential bifurcation steps when we do allow for interactions. A comparison of these two figures and the two lists of important factors in Table 3 shows that, in our case study, we find the *same shortlist*. The individual values, however, do differ: *interactions are important*.

In the next section, we apply sequential bifurcation to the other two supply chain configurations (the Current and Next generations) and we interpret these sequential bifurcation results through our knowledge of the simulated real system.

## 3 Case Study: Ericsson's Supply Chain Simulation Models

We first discuss some programming issues and then we define the inputs and outputs of the three simulation models. Finally, we present the results.

### *3.1 Programming the Three Simulation Models*

We give only a short description of the three supply chain configurations and their simulation models; for details we refer to Persson and Olhager (2002). At the start of our sequential bifurcation, we have three simulation models programmed in the Taylor II simulation software for discrete event simulations; see Incontrol (2003). We conduct our sequential bifurcation via Microsoft Excel, using the batch run mode in Taylor II. We store input–output data in Excel worksheets. This set-up facilitates the analysis of the simulation input–output data, but it constrains the set-up of the experiment. For instance, we cannot control the pseudorandom numbers in the batch mode of Taylor II. Hence, we cannot apply common pseudorandom numbers nor can we guarantee absence of overlap in the pseudorandom numbers; we conjecture that the probability of overlap is negligible in practice.

In order to validate the simulation models, Persson and Olhager (2002) examined the simulation of the Current supply chain configuration. They were able to validate this model because data were available at that time from the real-world supply chain. More precisely, they validated the model of the Current configuration through a structured walkthrough with the engineers and managers who are familiar with the system being modeled (see, also, Balci, 2001). Next, they developed the models for the other two supply chains from the Current model. They validated the model of the Old supply chain configuration through a less thorough walkthrough. The model of the Next Generation supply chain was not validated at all: Ericsson deemed this acceptable because this model was built from a validated Current model and this supply chain did not even exist at that time.

## 3.2 Inputs and Outputs of the Three Simulation Models

For our case study, we now apply sequential bifurcation assuming a metamodel consisting of a first-order polynomial augmented with two-factor interactions. This metamodel requires a foldover design, as described in Section 2.4. At the start of the procedure, we simulate the two extreme combinations which themselves form a foldover design. Then, as in Section 2.3, in the second step, we set the first 79 factors to their individual high levels and the remaining factors to their low levels; all these levels are reversed in the mirror image scenario.

Note that the number of simulation observations would be much smaller if we were to assume a first-order polynomial (see Figure 3), as this assumption avoids the mirror scenarios required by the foldover principle. Also, this assumption may lead to a different path through the list of individual factors, that is, a different sequence of simulated scenarios (compare Figures 3 and 4). In our case study, we felt almost certain that interactions would be important, so Figure 3 is meant only to explain the basics of sequential bifurcation.

The environmental factors (labeled 80 through 92 in Section 2) are the demand for products, the process yield, and the percentage of scrap at each test station. It can be proved that creating one group of environmental factors and one group of controllable factors enables estimation of sums of interactions between individual controllable and environmental factors; see, for example, Lewis and Dean (2001).

We label the factors such that all factors have the same meaning in the three simulation models. To achieve this, we introduce *dummy factors* for the Current and the Next Generation models that represent those factors that are removed as the supply chain is changed. Such dummy factors have zero effects but simplify the calculations and interpretations of the sequential bifurcation results.

Gunn and Nahavandi (2002) showed that initial conditions are important in manufacturing simulations. Therefore we use a warm-up period in this case study in order to be able to assume that the outputs of interest are in a *steady state*; see Persson and Olhager (2002). We determine this period by applying Welch's procedure, described by Law and Kelton (2000). This procedure gives a warm-up period of four weeks. (The warm-up period used by Persson and Olhager was only one week; they determined this period through a rule-of-thumb, namely, that the warm-up period should be three times the longest lead-time in the simulation.) After this warm-up period, we run each scenario for 16 weeks of production. This runlength seems to give steady-state output.

Each simulation model gives several outputs, but we focus on a single output, namely, the average weekly cost of the total supply chain in the steady state. This cost is calculated from inventory-related costs and quality-related costs. The inventory-related costs are calculated from the inventory levels throughout the supply chain and the current value of each product at each step in the chain. The quality-related costs concern yield, scrap, and modification time multiplied by the modification cost. The rework affects the inventory-related costs with higher levels of inventory as the reworked products once again flow through the supply chain.

Different outputs may have different important factors so the sequential bifurcation paths may differ.

## 3.3 Results for Three Simulation Models

The *aggregated effects* of the Old supply chain exceed those of the Next Generation supply chain, because the former aggregates more (positive) individual effects. For example, the Current simulation model has 14 dummy factors (which have zero effects), so the first sequential bifurcation step gives a smaller main (group) effect: for the Current model this effect is 7,101,983, whereas it is 15,016,102 for the Old model.

Furthermore, the *shortlists* are slightly shorter for the Current and the Next Generation models. The individual factors on the three shortlists are the same, except that the Next Generation model includes on its shortlist the extra factors 91 (product demand), 44, and 46 (where the latter two factors represent transportation between operations). The most important factor (92) is the demand for one of Ericsson's fast-selling products. The other factors represent transportation and yield.

Figure 5 illustrates how the estimated *upper limits U* for main effects decrease as new observations are obtained. Furthermore, this figure shows that, for the Old supply chain, the most important individual main effect, that of factor 92, has already been identified and estimated after only ten steps. The next important factor, 49, is identified after 16 observations, and so on.

In order to *verify* the shortlist produced by the sequential bifurcation, we make some confirmatory observations and test the effects of the remaining "unimportant"

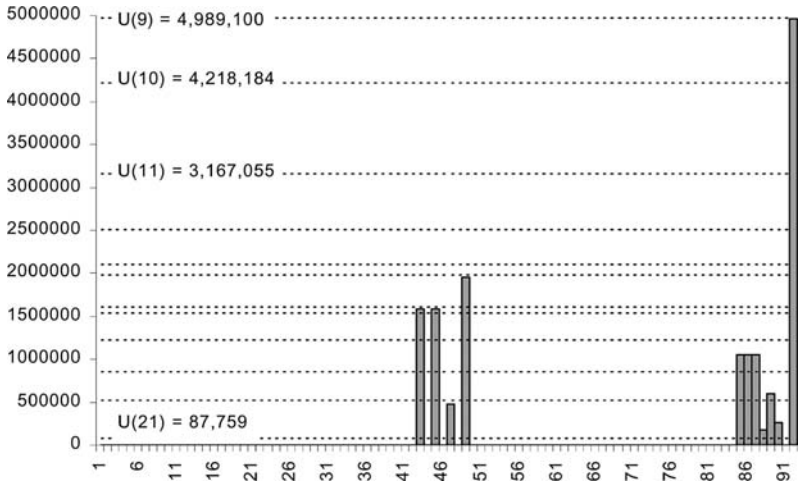



FIGURE 5. Upper limit $U(i)$ after step $i (i = 9, \ldots, 21)$ and individual main effect estimates (shaded bars) versus the factor label $j$ ($j = 1, \ldots, 92$) in the Old supply-chain simulation.

factors in the Current model. First, we set all the *unimportant factors* at their low values, while keeping the important factors fixed at their base or nominal values, which are coded as zero. Second, we switch all the unimportant factors to their high values, while keeping the important factors fixed. (These two scenarios are not used in sequential bifurcation to reach the shortlist.) We again replicate these scenarios five times. These replicates allow averages, standard deviations, and a $t$-statistic to be calculated. The resulting test statistic is nonsignificant and we conclude that the sequential bifurcation shortlist is valid. Persson (2003) gives many more details on both the case study and the application of sequential bifurcation to create the shortlists.

## 4 Discussion

The technique of sequential bifurcation is an important and useful method for identifying important factors in experiments with simulation models that involve a large number of factors. We have demonstrated the steps of this technique through a case study on three supply chain configurations in the Swedish mobile communications industry. We have formalized the assumptions of the technique and found that, in practice, these assumptions may not be too restrictive, as our case study illustrates. We have extended the technique of sequential bifurcation to random (as opposed to deterministic) simulations.

In a companion paper (Kleijnen et al., 2003), we changed the metamodel in (1) after the screening phase, as follows. For those controllable factors found to be important by sequential bifurcation, we augmented (1) with *quadratic* effects to form a predictive model for optimization. For those environmental or noise factors identified by sequential bifurcation as important, we created environmental scenarios through Latin hypercube sampling for robustness analysis.

Further research is needed to derive the overall probability of correctly classifying the individual factors as important or unimportant in our sequential procedure, which tests each factor group individually; see Lewis and Dean (2001, pages 663–664), Nelson (2003), Westfall et al. (1998), and Wan et al. (2004). Also, the extension of sequential bifurcation from single to multiple responses is an important practical and theoretical problem that requires further work.

*Acknowledgement.* This research was sponsored by KetenLogistiek en Informatie en CommunicatieTechnologie/Interdepartementale Commissie voor Economische Structuurversterking ("Chain networks, Clusters and ICT").

## *References*

Andres, T. H. and Hajas, W.C. (1993). Using iterated fractional factorial design to screen parameters in sensitivity analysis of a probabilistic risk assessment model. *Proceedings of the Joint International Conference on Mathematical Models and Supercomputing in*

*Nuclear Applications*, 2. Karlsruhe, Germany, 19–23 April 1993. Editors: H. Küters, E. Stein and W. Werner, pages 328–337.

Balci, O. (2001). A methodology for accreditation of modeling and simulation applications. *Transactions on Modeling and Computer Simulation*, **11**, 352–377.

Banks, J., Buckley, S., Jain, S., and Lendermann, P. (2002). Panel session: Opportunities for simulation in supply chain management. *Proceedings of the 2002 Winter Simulation Conference*. Editors: E. Yücesan, C.H. Chen, J.L. Snowdon, and J.M. Charnes, pages 1652–1658.

Bettonvil, B. (1990). *Detection of Important Factors by Sequential Bifurcation*. Tilburg University Press, Tilburg.

Bettonvil, B. and Kleijnen, J.P.C. (1996). Searching for important factors in simulation models with many factors: Sequential bifurcation. *European Journal of Operational Research*, **96**, 180–194.

Box, G.E.P. and Wilson, K.B. (1951). On the experimental attainment of optimum conditions. *Journal Royal Statistical Society B*, **13**, 1–38.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building*. Wiley, New York.

Campolongo, F., Kleijnen, J.P.C., and Andres, T. (2000). Screening methods. In *Sensitivity Analysis*. Editors: A. Saltelli, K. Chan, and E.M. Scott, pages 65–89. Wiley, Chichester, England.

Cheng, R.C.H. (1997). Searching for important factors: Sequential bifurcation under uncertainty. *Proceedings of the 1997 Winter Simulation Conference*. Editors: S. Andradottir, K.J. Healy, D.H. Withers, and B.L. Nelson, pages 275–280.

Cheng, R.C.H. and Holland, W. (1999). Stochastic sequential bifurcation: Practical issues. *Proceedings of UK Simulation Society Conference*. Editors: D. Al-Dabass and R.C.H. Cheng, UKSIM, Nottingham, pages 74–80.

Dean, A. M. and Voss, D. T. (1999). *Design and Analysis of Experiments*, Springer, New York.

De Vos, C., Saatkamp, H.W., Nielen, M., and Huirne, R.B.M. (2003). Sensitivity analysis to evaluate the impact of uncertain factors in a scenario tree model for classical swine fever introduction. Working Paper, Wageningen University, The Netherlands.

Gunn, B. and Nahavandi, S. (2002). Why initial conditions are important. *Proceedings of the 2002 Winter Simulation Conference*. Editors: E. Yücesan, C.H. Chen, J.L. Snowdon, and J.M. Charnes, pages 557–562.

Incontrol (2003). Incontrol Enterprise Dynamics. http://www.EnterpriseDynamics.com.

Kleijnen, J.P.C. (1987). *Statistical Tools for Simulation Practitioners*. Marcel Dekker, New York.

Kleijnen, J.P.C. (2005). Supply chain simulation tools and techniques: A survey. *International Journal of Simulation and Process Modelling*, **1**(1/2), 82–89.

Kleijnen, J.P.C. and Sargent, R.G. (2000). A methodology for the fitting and validation of metamodels in simulation. *European Journal of Operational Research*, **120**, 14–29.

Kleijnen, J.P.C., Bettonvil, B., and Persson, F. (2003). Robust solutions for supply chain management: Simulation, optimization, and risk analysis. http://center.kub.nl/staff/kleijnen/papers.html.

Kleijnen, J.P.C., Sanchez, S.M., Lucas, T.W., and Cioppa, T.M. (2005). A user's guide to the brave new world of designing simulation experiments. *INFORMS Journal on Computing*, **17**, 263–289.

Law, A.M. and Kelton, W.D. (2000). *Simulation Modeling and Analysis*, third edition. McGraw-Hill, Boston.

Lee, H.L., Padmanabhan, V., and Whang, S. (1997). The bullwhip effect in supply chains. *Sloan Management Review*, **38**, Spring, 93–102.

Lewis, S.M. and Dean, A.M. (2001). Detection of interactions in experiments on large numbers of factors (with discussion). *Journal Royal Statistical Society B*, **63**, 633–672.

Morris, M. D. (1991). Factorial plans for preliminary computational experiments. *Technometrics*, **33**, 161–174.

Myers, R.H. and Montgomery, D.C. (2002). *Response Surface Methodology: Process and Product Optimization using Designed Experiments*, second edition. John Wiley and Sons, New York.

Nelson, B. L. (2003). Private communication (nelsonb@primal.iems.nwu.edu).

Persson, F. (2003). Discrete event simulation of supply chains: Modelling, validation, and analysis. Doctoral dissertation. Profil 20, Linköping Institute of Technology.

Persson, F. and Olhager, J. (2002). Performance simulation of supply chain designs. *International Journal of Production Economics*, **77**, 231–245.

Simchi-Levi, D., Kaminsky, P., and Simchi-Levi, E. (2000). *Designing and Managing the Supply Chain. Concepts, Strategies, and Case Studies*. Irwin/McGraw-Hill, Boston.

Wan, H., Ankenman, B. E., and Nelson, B. L. (2004). Controlled sequential bifurcation: A new factor-Screening method for discrete-event simulation. IEMS Technical Report 04-011, Northwestern University. http://www.iems.northwestern.edu/content/papers.asp

Westfall, P.H., Young, S.S., and Lin, D.K.J. (1998). Forward selection error control in the analysis of supersaturated designs. *Statistica Sinica*, **8**, 101–117.

Wu, C.F.J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley and Sons, New York.

Zeigler, B.P., Praehofer, H., and Kim, T.G. (2000). *Theory of Modelling and Simulation*, second edition. Academic Press, San Diego.

# 14
# Screening the Input Variables to a Computer Model Via Analysis of Variance and Visualization

Matthias Schonlau and William J. Welch

An experiment involving a complex computer model or code may have tens or even hundreds of input variables and, hence, the identification of the more important variables (screening) is often crucial. Methods are described for decomposing a complex input–output relationship into effects. Effects are more easily understood because each is due to only one or a small number of input variables. They can be assessed for importance either visually or via a functional analysis of variance. Effects are estimated from flexible approximations to the input–output relationships of the computer model. This allows complex nonlinear and interaction relationships to be identified. The methodology is demonstrated on a computer model of the relationship between environmental policy and the world economy.

## 1 Introduction

Computer models, also known as "math models" or "codes", are now frequently used in engineering, science, and many other disciplines. To run the computer model software, the experimenter provides quantitative values for various input (explanatory) variables. The code then computes values for one or more output (response) variables. For instance, in a model of Arctic sea ice (Chapman et al., 1994), the input variables included rate of snowfall and ice albedo and the code produced values of ice mass, and so on. In circuit-design models (see, for example, Aslett et al., 1998), the input variables are transistor widths and other engineering parameters, and the output variables are measures of circuit performance such as time delays.

Often, the computer model will be expensive to run, for example, if it solves a large number of differential equations that may require several hours or more of computer time. Thus, in a computer experiment, that is, an experiment with several runs of the computer model, there is need for careful design or choice of the values of the input variables and careful analysis of the data produced.

One major difference from traditional design and analysis of experiments with physical measurements is that computer models are often deterministic. Two runs of the code with the same set of values for the input variables would give identical results across the two runs for each output variable. Nonetheless, there will often

be considerable uncertainty in drawing conclusions about the behavior of the input–output relationships from a limited number of runs, and statistical methods are required to characterize the uncertainty. The management of uncertainty is especially critical when the computer model has a high-dimensional set of input variables.

In applications such as the Arctic sea ice model (Chapman et al., 1994) mentioned above, a strategic objective of a preliminary computer experiment is screening: finding the important input variables. Screening is not a trivial task because the computer model is typically complex, and the relationships between input variables and output variables are not obvious. A common approach is to approximate the relationship by a statistical surrogate model, which is easier to explore. This is particularly useful when there are many input variables.

An example, which is discussed in this chapter, is the "Wonderland" model of Lempert et al. (2003), adapted from Herbert and Leeves (1998). In this case study, 41 input variables are manipulated, relating to population growth, economic activity, changes in environmental conditions, and other economic and demographic variables. The output is a quasi global human development index (HDI) which is a weighted index of net output per capita, death rates, annual flow of pollution, and the carrying capacity of the environment, spanning both "northern" and "southern" countries. The model has many output variables under various policy assumptions; we consider only one, corresponding to a "limits to growth" policy. Under this scenario, economic growth is intentionally limited by a constraint on global emissions. After 2010, both hemispheres must set carbon taxes high enough to achieve zero growth in emissions levels. Larger values of HDI correspond to greater human development and are better; see Lempert et al. (2003) for a full description of this measure.

Figure 1 shows scatter plots of the raw data from a Latin hypercube experimental design (see McKay et al., 1979) with 500 runs of the Wonderland code. The output variable HDI is plotted against two of the input variables shown in Section 6 to be important: economic innovation in the north (e.inov.n) and sustainable pollution in the south (v.spoll.s).

The first plot suggests a slight upward trend in HDI with e.inov.n. It is shown in Section 6 that the trend is actually very strong; it looks weak here because there is considerable masking from other variables. The second plot shows a very rapid dropoff in HDI for low values of v.spoll.s. This nonlinearity in the computer model would have gone unnoticed without the three points on the left with the lowest HDI values. Thus, a design with fewer runs, or with fewer levels of each input variable, may well have missed the region containing these three points. (Note also that very small values of v.spoll.s do not always give such extreme values of HDI.) In our experience, nonlinear effects are common in computer experiments because the input variables often cover wide ranges. We explore the Wonderland application further in Section 6, but it is hoped we have already illustrated some of the potential difficulties in screening the input variables of a computer model: large dimensionality, complex nonlinearities, and masking.

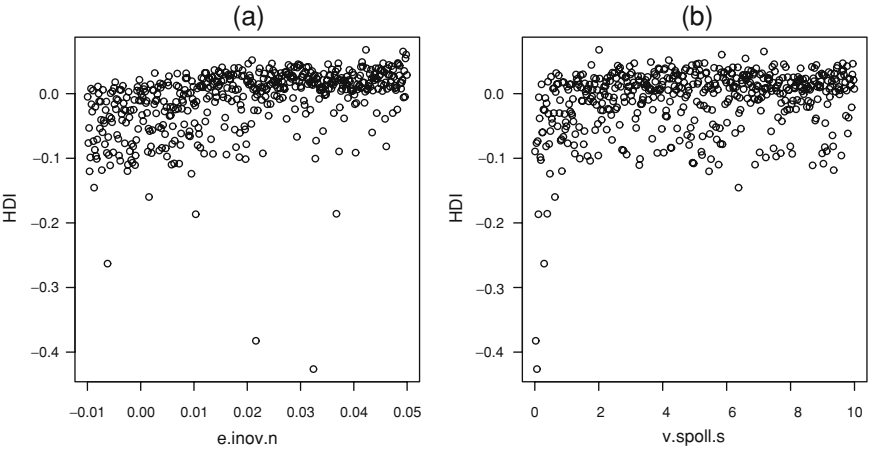



FIGURE 1. Human development index (HDI) from the Wonderland model plotted against (a) economic innovation in the north (e.inov.n) and against (b) sustainable pollution in the south (v.spoll.s).

There is a spectrum of methods proposed for screening variables in a computer experiment. They differ mainly in the assumptions they make about the form of an input–output relationship: with stronger assumptions, fewer runs are typically required.

Iman and Conover (1980) built a rank-regression approximation of a computer model of the discharge of a nuclear isotope from radioactive waste. With seven input variables, they used 200 runs in a Latin hypercube design. A sensitivity analysis followed from the least squares estimates of the coefficients in the first-order rank-regression model. (A sensitivity analysis, which explores how sensitive the output variable is to changes in the input variables, is similar to screening.) Morris (1991) described a screening method where the number of runs is a small multiple of the number of input variables. The method does not attempt to model the input–output relationship(s) of a computer code, rather it attempts to divide the variables qualitatively into three categories: unimportant; those with linear and additive effects; and those with more complex, nonlinear, or interaction effects. Twenty variables in a heat-transfer model were investigated using 108 runs.

With even fewer runs relative to the number of input variables, Bettonvil and Kleijnen (1996) used a sequential bifurcation algorithm (see Chapter 13) to analyze a large deterministic global-climate model. The output is the worldwide $CO_2$ circulation in the year 2100. The model has 281 input variables, 15 of which were identified as important after 154 runs. The sequential bifurcation algorithm makes several strong assumptions to enable an experiment with fewer runs than input variables (a supersaturated design—see Chapter 8). Each variable is considered at only two levels, and effects are assumed to be linear and additive. Moreover, the direction (sign) of each effect must be known a priori. The sequential bifurcation

method was followed up with a traditional Resolution IV design (Chapter 1) for the most important factors in order to estimate a response surface model.

Gu and Wahba (1993) used a smoothing-spline approach with some similarities to the method described in this chapter, albeit in a context where random error is present. They approximated main effects and some specified two-variable interaction effects by spline functions. Their example had only three explanatory variables, so screening was not an issue. Nonetheless, their approach parallels the methodology we describe in this chapter, with a decomposition of a function into effects due to small numbers of variables, visualization of the effects, and an analysis of variance (ANOVA) decomposition of the total function variability.

The approach to screening the input variables in a computer model described in this chapter is based on a Gaussian random-function approximator to an input–output relationship. This is a flexible, data-adaptive paradigm with a long history in the analysis of computer experiments. Similarly, decomposing the random-function approximator into low-order effects for the purposes of identifying and examining the important effects has been in use for some time. The estimated effects are visualized or quantified via a functional ANOVA decomposition. In an experiment with six input factors and 32 runs, Sacks et al. (1989) identified the important (nonlinear) main effects and two-variable interaction effects. Welch et al. (1992) described a stepwise method for adding important input variables to the statistical approximator and visualized the important effects. They were able to find the important nonlinear and interaction effects among 20 input variables with 50 runs. Chapman et al. (1994) and Gough and Welch (1994) performed sensitivity analyses of climate models with 13 and 7 input variables, respectively, and Mrawira et al. (1999) were able to deal with 35 input variables in a civil-engineering application. With up to 36 variables, Aslett et al. (1998) and Bernardo et al. (1992) used visualization of important effects to guide the sequential optimization of electronic circuit designs. Santner et al. (2003, Chapter 7) also summarized this approach.

Thus, decomposition of a random-function approximator of a computer model into low-dimensional effects, in order to identify the important effects and examine them visually and quantitatively, has been widely applied and reported by many authors. However, the implementation of these methods has not been described, with the partial exception of Schonlau (1997), a shortcoming that we address in this chapter.

The chapter is organized as follows. Section 2 reviews the key underlying random-function approximator. Effects are defined in Section 3, leading to a functional ANOVA, and Section 4 describes their estimation. Section 5 summarizes the steps in the work flow for identifying and visualizing the important estimated effects. This approach has proved to be a powerful screening tool, as evidenced by the above examples. In Section 6, we return to the Wonderland model and demonstrate how the methodology is used. Some concluding remarks are given in Section 7. Some details of the derivation of the best linear unbiased predictor of an effect are provided in Appendix A, and Appendix B shows how the high-dimensional integrals required for the estimated effects, for their pointwise standard errors, and

for the ANOVA decomposition boil down to a series of low-dimensional integrals under certain, fairly common, conditions.

## 2 The Random-Function Approximator

Here, we give a brief review of methods for the analysis of computer experiments, concentrating on statistical approximation of the computer model. Strategies for the design and analysis of computer experiments have been described by many authors, including Currin et al. (1991), Koehler and Owen (1996), Sacks et al. (1989), Santner et al. (2003), and Welch et al. (1992). All these authors take into account the deterministic nature of a code, such as the Wonderland model, and also provide uncertainty measures via a statistical approximation model.

In general, suppose that a code is run $n$ times in a computer experiment. Each run has a different set of values for the $d$-dimensional vector of input variables, $\boldsymbol{x} = (x_1, \ldots, x_d)^T$. A particular output variable is denoted by $y(\boldsymbol{x})$. With several output variables, each is treated separately. The data consist of $n$ input vectors, $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$, chosen from an input region of interest, $\chi$, and the vector of $n$ corresponding output values, denoted by $\boldsymbol{y}$.

Following the approach of the above authors, the output variable $y(\boldsymbol{x})$ is treated as a realization of a random function:

$$Y(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), \ldots, f_h(\boldsymbol{x})]'$ is a vector of $h$ known regression functions, $'$ denotes transpose, $\boldsymbol{\beta}$ is a $h \times 1$ vector of parameters to be estimated, and $Z$ is a Gaussian stochastic process indexed by $\boldsymbol{x}$. It is assumed that $Z(\boldsymbol{x})$ has mean zero and constant variance, $\sigma^2$, for all $\boldsymbol{x}$. The covariance between $Z(\boldsymbol{x})$ and $Z(\tilde{\boldsymbol{x}})$ at two input vectors, $\boldsymbol{x} = (x_1, \ldots, x_d)'$ and $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_d)'$, is

$$\mathrm{Cov}[Z(\boldsymbol{x}), Z(\tilde{\boldsymbol{x}})] = \sigma^2 R(\boldsymbol{x}, \tilde{\boldsymbol{x}}), \tag{2}$$

where $R(\cdot, \cdot)$ is a "correlation function" and $\tilde{\boldsymbol{x}}$ denotes a different set of input values from $\boldsymbol{x}$.

The correlation function $R(\cdot, \cdot)$ in (2) is central to this statistical model. The *power-exponential* class of correlation functions is a popular choice, for its computational simplicity and because it has been successful in many applications. The power-exponential correlation function is

$$R(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \prod_{j=1}^{d} \exp(-\theta_j |x_j - \tilde{x}_j|^{p_j}), \tag{3}$$

where $\theta_j \geq 0$ and $0 < p_j \leq 2 (j = 1, \ldots, d)$ are parameters that can be estimated from the data, often via maximum likelihood. The $p_j$ can be interpreted as smoothness parameters—the output surface is smoother with respect to $x_j$ as $p_j$ increases. For $p_j = 2$, the surface is infinitely differentiable. For $0 < p_j < 2$, the surface is continuous, but not differentiable. As $p_j$ increases between 0 and 2, however, the

surface appears to fluctuate less and less, and in this sense could be said to be smoother. The $\theta_j$ indicate the extent to which the variation in the output function is local with respect to $x_j$. If $\theta_j$ is large, the correlation (3) between observations or outputs at $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ falls rapidly with the distance between $x_j$ and $\tilde{x}_j$, and the function is difficult to predict in the $x_j$ direction.

We next describe the first steps in the derivation of the best linear unbiased predictor (BLUP) of $Y(\boldsymbol{x})$ at an untried input vector $\boldsymbol{x}$ (see, for example, Sacks et al., 1989). Similar steps are used in Section 4 to estimate the effects of one, two, or more input variables. It is then apparent how to adapt results and computational methods for predicting $Y(\boldsymbol{x})$ to the problem of estimating such effects.

Following the random-function model (1), consider the prediction of $Y(\boldsymbol{x})$ by $\hat{Y}(\boldsymbol{x}) = \boldsymbol{a}'(\boldsymbol{x})\boldsymbol{y}$, that is, a linear combination of the $n$ values of the output variable observed in the experiment. The best linear unbiased predictor is obtained by minimizing the mean squared error of the linear predictor or approximator, $\hat{Y}(\boldsymbol{x})$. The mean squared error, MSE$[\hat{Y}(\boldsymbol{x})]$, is

$$\begin{aligned}\mathrm{E}[Y(\boldsymbol{x}) - \hat{Y}(\boldsymbol{x})]^2 &= \mathrm{E}[\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}) - \boldsymbol{a}'(\boldsymbol{x})(\boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{z})]^2 \\ &= \{[\boldsymbol{f}'(\boldsymbol{x}) - \boldsymbol{a}'(\boldsymbol{x})\boldsymbol{F}]\boldsymbol{\beta}\}^2 \\ &\quad + \mathrm{Var}[Z(\boldsymbol{x})] + \boldsymbol{a}'(\boldsymbol{x})\mathrm{Cov}(\boldsymbol{z})\boldsymbol{a}(\boldsymbol{x}) - 2\boldsymbol{a}'(\boldsymbol{x})\mathrm{Cov}[Z(\boldsymbol{x}), \boldsymbol{z}],\end{aligned}$$

where $F$ is the $n \times h$ matrix with row $i$ containing the regression functions $\boldsymbol{f}'(\boldsymbol{x}^{(i)})$ for run $i$ in the experimental plan, and $\boldsymbol{z} = [Z(\boldsymbol{x}^{(1)}), \ldots, Z(\boldsymbol{x}^{(n)})]'$ is the $n \times 1$ vector of random $Z$ values, with element $i$ corresponding to run $i$. From the covariance function (2) we can write Cov$(\boldsymbol{z})$ as $\sigma^2\boldsymbol{R}$, where $\boldsymbol{R}$ is an $n \times n$ matrix with element $(i, j)$ given $R(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$, and Cov$[Z(\boldsymbol{x}), \boldsymbol{z})]$ as $\sigma^2\boldsymbol{r}(\boldsymbol{x})$, where $\boldsymbol{r}(\boldsymbol{x})$ is an $n \times 1$ vector with element $i$ given by $R(\boldsymbol{x}, \boldsymbol{x}^{(i)})$. With this notation, the mean squared error of $\hat{Y}(\boldsymbol{x})$ is

$$\begin{aligned}\mathrm{MSE}[\hat{Y}(\boldsymbol{x})] &= \{[\boldsymbol{f}'(\boldsymbol{x}) - \boldsymbol{a}'(\boldsymbol{x})\boldsymbol{F}]\boldsymbol{\beta}\}^2 \\ &\quad + \mathrm{Var}[Z(\boldsymbol{x})] + \sigma^2\boldsymbol{a}'(\boldsymbol{x})\boldsymbol{R}\boldsymbol{a}(\boldsymbol{x}) - 2\sigma^2\boldsymbol{a}'(\boldsymbol{x})\boldsymbol{r}(\boldsymbol{x}). \qquad (4)\end{aligned}$$

Some further simplification of this expression is possible, for example, by using the fact that $\mathrm{Var}[Z(\boldsymbol{x})] = \sigma^2$, by assumption. We leave the mean squared error in this form, however, to facilitate comparison with its counterpart in Section 4 for the estimated effect of a group of variables.

We now choose $\boldsymbol{a}(\boldsymbol{x})$ to minimize (4). To avoid an unbounded contribution from the first term on the right-hand side of (4) from large elements in $\boldsymbol{\beta}$, the contribution is eliminated by imposing the constraint

$$\boldsymbol{F}\boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}).$$

This constraint is also sometimes motivated by unbiasedness, that is, from $\mathrm{E}[\hat{Y}(\boldsymbol{x})] = \mathrm{E}[Y(\boldsymbol{x})]$ for all $\boldsymbol{\beta}$. Thus, the best linear unbiased predictor, or optimal value of $\boldsymbol{a}(\boldsymbol{x})$, results from the following optimization problem,

$$\min_{a(\boldsymbol{x})} \mathrm{Var}[Z(\boldsymbol{x})] + \sigma^2\boldsymbol{a}'(\boldsymbol{x})\boldsymbol{R}\boldsymbol{a}(\boldsymbol{x}) - 2\sigma^2\boldsymbol{a}'(\boldsymbol{x})\boldsymbol{r}(\boldsymbol{x}) \qquad (5)$$

subject to

$$\boldsymbol{F}\boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}).$$

The optimal $\boldsymbol{a}(\boldsymbol{x})$ turns out to give the following form for the BLUP (or approximator) (see, for example, Sacks et al., 1989),

$$\hat{Y}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})\hat{\boldsymbol{\beta}} + r'(\boldsymbol{x})\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{F}\hat{\boldsymbol{\beta}}), \tag{6}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{F}'\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}\boldsymbol{R}^{-1}\boldsymbol{y}$ is the generalized least squares estimator of $\boldsymbol{\beta}$. If we put the optimal $\boldsymbol{a}(\boldsymbol{x})$ into the expression for the mean squared error (4), we obtain the following standard error, se$[\hat{Y}(\boldsymbol{x})]$, for $\hat{Y}(\boldsymbol{x})$:

$$\begin{aligned}\text{se}^2[\hat{Y}(\boldsymbol{x})] = \text{Var}[Z(\boldsymbol{x})] &- \sigma^2\boldsymbol{r}(\boldsymbol{x})'\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x}) \\ &+ \sigma^2[\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{F}'\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x})]'(\boldsymbol{F}'\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}[\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{F}'\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x})]. \end{aligned} \tag{7}$$

This formula ignores the uncertainty from estimating the correlation parameters, for example, the $\theta_j$ and $p_j$ in (3). Some comments on this issue are made in Section 7.

## 3 Effects

The important input variables are those that have large effects on the output variable. As with traditional analysis of variance, we can look at the main effects of single variables, or the joint or interaction effects of several variables at a time.

Suppose that we are interested in the effect of a subset of input variables, held in a vector $\boldsymbol{x}_e$, where $e$ denotes the set of subscripts of the variables of interest. The vector of remaining variables is denoted by $\boldsymbol{x}_{-e}$. For example, when interest is in the effects of $x_1$ and $x_2$ among $d > 2$ variables, we have $e = \{1, 2\}$ and $\boldsymbol{x}_e = (x_1, x_2)$, whereupon $\boldsymbol{x}_{-e} = (x_3, \ldots, x_d)$. Without loss of generality we rearrange the order of the input variables so that we can write $\boldsymbol{x} = (\boldsymbol{x}_e, \boldsymbol{x}_{-e})$. To obtain a unique and workable definition of the effect of $x_e$ is essentially the problem of how to deal with the variables in $\boldsymbol{x}_{-e}$. We next discuss several ways of approaching this problem.

Keeping the variables in $\boldsymbol{x}_{-e}$ fixed requires little new methodology. We consider $y(\boldsymbol{x}_e, \boldsymbol{x}_{-e})$ as a function of $\boldsymbol{x}_e$, with $\boldsymbol{x}_{-e}$ fixed, for example, at the variable midranges. Estimates and pointwise standard errors of such effects follow immediately from (6) and (7), with $\hat{Y}(\boldsymbol{x}_e, \boldsymbol{x}_{-e})$ and se$[\hat{Y}(\boldsymbol{x}_e, \boldsymbol{x}_{-e})]$ considered as functions of $x_e$. There are two serious disadvantages of this method, however. First, in the presence of interaction effects involving one or more variables in $\boldsymbol{x}_e$ and one or more variables in $\boldsymbol{x}_{-e}$, the magnitude of the effect of $\boldsymbol{x}_e$ may change depending on the levels chosen for $\boldsymbol{x}_{-e}$, and thus the effect of $\boldsymbol{x}_e$ is not isolated. Consequently, there is no straightforward decomposition of the total variation in $y(\boldsymbol{x})$, or its predictor $\hat{Y}(\boldsymbol{x})$, into contributions from various effects.

Alternatively, we may define an effect by "integrating out" the other variables. Under certain conditions, this leads to a simple decomposition of $y(\boldsymbol{x})$ into contributions from various effects, with a corresponding decomposition of the total

variance of $y(\boldsymbol{x})$ over $\chi$. Moreover, as we show in Section 4, these effects and their variance contributions can be easily estimated. Hence, defining an effect by integrating out the other variables is the method pursued for the remainder of this chapter.

For a convenient decomposition of $y(\boldsymbol{x})$, we need two conditions on the region of interest of the input variables. First, $\chi$ is assumed to be a direct product of one-dimensional regions, which we write as

$$\chi = \otimes_{j=1}^{d} \chi_j, \tag{8}$$

where $\chi_j$ denotes the values of interest for variable $x_j$, for instance, a continuous interval or a discrete set of points (for which integration is interpreted as summation). Second, we assume that integration is with respect to a weight function, $w(\boldsymbol{x})$, which is a product of functions of one input variable at a time:

$$w(\boldsymbol{x}) = \prod_{j=1}^{d} w_j(x_j) \quad \text{for} \quad x_j \in \chi_j,\ j = 1, \ldots, d. \tag{9}$$

Often, the weight function $w_j(x_j)$ for $x_j$ is chosen to be a uniform distribution, representing equal interest across the range of values for $x_j$. In other applications, $x_j$ might be a variable in the computer code because its value in nature is uncertain. If this uncertainty is represented by a given statistical distribution, for example, a normal distribution, then the distribution would be used as the weight function, $w_j(x_j)$. The conditions (8) and (9) occur frequently in applications; a minor relaxation of them is discussed in Section 4.

Under the assumptions (8) and (9), the *marginal effect*, $\bar{y}_e(\boldsymbol{x}_e)$, of $\boldsymbol{x}_e$ is defined by integrating out the other variables,

$$\bar{y}_e(\boldsymbol{x}_e) = \int_{\otimes_{j \notin e} \chi_j} y(\boldsymbol{x}_e, \boldsymbol{x}_{-e}) \prod_{j \notin e} w_j(x_j) dx_j \quad \text{for } \boldsymbol{x}_e \in \otimes_{j \in e} \chi_j. \tag{10}$$

Note that a marginal effect is the overall effect of all the variables in $\boldsymbol{x}_e$. With just one variable in $\boldsymbol{x}_e$, we call this a *main* effect; with two or more variables, we call this a *joint* effect.

We use the marginal effects (10) to decompose $y(\boldsymbol{x})$ as follows into corrected or adjusted effects involving no variables, one variable at a time, two variables at a time, and so on, up to the contribution from all the variables:

$$y(\boldsymbol{x}) = \mu_0 + \sum_{j=1}^{d} + \mu_j(x_j) + \sum_{j=1}^{d-1} \sum_{j'=j+1}^{d} \mu_{jj'}(x_j, x_{j'}) + \cdots + \mu_{1\ldots d}(x_1, \ldots, x_d) \tag{11}$$

for $\boldsymbol{x} \in \chi$, where

$$\mu_0 = \int_{\chi} y(\boldsymbol{x}) w(\boldsymbol{x}) d\boldsymbol{x}$$

is an overall average,

$$\mu_j(x_j) = \bar{y}_j(x_j) - \mu_0 \quad \text{for } x_j \in \chi_j \tag{12}$$

is the corrected main effect of $x_j$,

$$\mu_{jj'}(x_j, x_{j'}) = \bar{y}_{jj'}(x_j, x_{j'}) - \mu_j(x_j) - \mu_{j'}(x_{j'}) - \mu_0 \quad \text{for } x_j, x_{j'} \in \chi_j \otimes \chi_{j'} \tag{13}$$

is the corrected joint effect or interaction effect of $x_j$ and $x_{j'}$, and so on. Thus, each corrected effect is the corresponding marginal effect corrected for all lower-order terms.

For example, suppose interest centers on the variables $x_1$ and $x_2$. If their interaction effect, $\mu_{12}(x_1, x_2)$, has an important magnitude, it is not meaningful to consider the effects of $x_1$ or $x_2$ in isolation. We would look at their overall joint effect,

$$\bar{y}_{12}(x_1, x_2) = \mu_0 + \mu_1(x_1) + \mu_2(x_2) + \mu_{12}(x_1, x_2) \text{ for } x_1, x_2 \in \chi_1 \otimes \chi_2.$$

Similar comments apply to higher-order effects. In practice, we will have to estimate the marginal effects, and hence the corrected effects, to decide which are important.

The effects (11) are orthogonal with respect to the weight function $w(\boldsymbol{x})$, leading to a decomposition of the total variance of $y(\boldsymbol{x})$, called the *ANOVA decomposition* or *functional analysis of variance* as follows,

$$\begin{aligned}\int_{\chi} [y(\boldsymbol{x}) - \mu_0]^2 w(\boldsymbol{x}) d\boldsymbol{x} = &\sum_{j=1}^{d} \int_{\chi_j} \mu_j^2(x_j) w_j(x_j) dx_j \\ &+ \sum_{j=1}^{d-1} \sum_{j'=j+1}^{d} \int_{\chi_j \otimes \chi_{j'}} \mu_{jj'}^2(x_j, x_{j'}) w_j(x_j) w_{j'}(x_{j'}) dx_j dx_{j'} \\ &+ \cdots + \int_{\chi} \mu_{1\ldots d}^2(x_1, \ldots, x_d) \prod_{j=1}^{d} w_j(x_j) dx_j. \end{aligned} \tag{14}$$

A quantitative measure of the importance of any effect, and hence the associated variables, follows from the percentage contribution of each term on the right-hand side to the total variance on the left. The functional analysis of variance (ANOVA) in (14) goes back at least as far as Hoeffding (1948).

## 4 Estimating the Effects

Estimating the marginal (main or joint) effects $\bar{y}_e(\boldsymbol{x}_e)$ in (10) is key to our approach for assessing the importance of variables. From the estimated marginal effects, we can also estimate the corrected effects in (11) and the ANOVA decomposition (14). Furthermore, when visualizing the large estimated effects it is easier to interpret main or joint effects than their corrected counterparts.

If $y(\boldsymbol{x})$ is treated as if it is a realization of the random function $Y(\boldsymbol{x})$ in (1), it follows that $\bar{y}_e(\boldsymbol{x}_e)$ is a realization of the analogously integrated random function,

$$\bar{Y}_e(\boldsymbol{x}_e) = \bar{\boldsymbol{f}}'_e(\boldsymbol{x}_e)\beta + \bar{Z}_e(x_e) \qquad \text{for } \boldsymbol{x}_e \in \otimes_{j\in e}\chi_j. \tag{15}$$

Here, $\bar{\boldsymbol{f}}_e(\boldsymbol{x}_e)$ and $\bar{Z}_e(\boldsymbol{x}_e)$ have the input variables not in $\boldsymbol{x}_e$ integrated out as in (10):

$$\bar{\boldsymbol{f}}_e(\boldsymbol{x}_e) = \int_{\otimes_{j\notin e}\chi_j} f(\boldsymbol{x}_e, \boldsymbol{x}_{-e}) \prod_{j\notin e} w_j(x_j)dx_j \qquad \text{for } \boldsymbol{x}_e \in \otimes_{j\in e}\chi_j \tag{16}$$

and

$$\bar{Z}_e(\boldsymbol{x}_e) = \int_{\otimes_{j\notin e}\chi_j} Z(\boldsymbol{x}_e, \boldsymbol{x}_{-e}) \prod_{j\notin e} w_j(x_j)dx_j \qquad \text{for } \boldsymbol{x}_e \in \otimes_{j\in e}\chi_j. \tag{17}$$

The statistical properties of the stochastic process $\bar{Z}_e(\boldsymbol{x}_e)$ and the derivation of the BLUP of $\bar{Y}_e(\boldsymbol{x}_e)$ are derived in Appendix A. It is shown that the BLUP of $\bar{Y}_e(\boldsymbol{x}_e)$ is

$$\hat{\bar{Y}}_e(\boldsymbol{x}_e) = \bar{\boldsymbol{f}}_e(\boldsymbol{x}_e)\hat{\boldsymbol{\beta}} + \bar{\boldsymbol{r}}'_e(\boldsymbol{x}_e)\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{F}\hat{\beta}) \tag{18}$$

and its standard error is given by

$$\begin{aligned} \mathrm{se}^2[\hat{\bar{Y}}_e(\boldsymbol{x}_e)] = {} & \mathrm{Var}[\bar{Z}_e(\boldsymbol{x}_e)] - \sigma^2\bar{\boldsymbol{r}}_e(\boldsymbol{x})'\boldsymbol{R}^{-1}\bar{\boldsymbol{r}}_e(\boldsymbol{x}_e) \\ & + \sigma^2[\bar{\boldsymbol{f}}_e(\boldsymbol{x}_e) - \boldsymbol{F}'\boldsymbol{R}^{-1}\bar{\boldsymbol{r}}_e(\boldsymbol{x}_e)]'(\boldsymbol{F}'\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}[\bar{\boldsymbol{f}}_e(\boldsymbol{x}_e) - \boldsymbol{F}'\boldsymbol{R}^{-1}\bar{\boldsymbol{r}}_e(\boldsymbol{x}_e)], \end{aligned} \tag{19}$$

where $\bar{\boldsymbol{r}}_e(\boldsymbol{x}_e)$ is defined following (22) in Appendix A.

In other words, software for computing the BLUP of $Y(x)$ and its standard error is easily modified for estimating effects and, hence, the ANOVA decomposition, provided that we can compute $\mathrm{Var}[\bar{Z}_e(\boldsymbol{x}_e)]$ from (21) in Appendix A, $\bar{\boldsymbol{r}}_e(\boldsymbol{x}_e)$ following (22), and $\bar{\boldsymbol{f}}_e(\boldsymbol{x}_e)$ in (16), quantities that will involve high-dimensional integrals in high-dimensional problems. These computations are described in Appendix B.

It is possible to relax the product-region condition (8) in some experiments. For example, Mrawira et al. (1999) dealt with several groups of variables where there were constraints like $x_1 \le x_2$. The triangular input space for such a group had a product arrangement with all other variables or groups of variables. Thus, in all the above formulas for estimated effects or their standard errors, we merely treat the variables in a group together as if they were a single variable. This means, however, that the estimated effect for a group cannot be decomposed further into contributions from its constituent variables.

## 5 Steps for Identifying and Visualizing the Important Estimated Effects

To screen the input variables, we carry out the following steps.

1. Estimate by maximum likelihood the unknown parameters, $\beta$ in (1), $\sigma^2$ in (2), and the correlation parameters, for example, the $\theta_j$ and $p_j$ in (3).

2. Before continuing with a screening analysis, it is prudent to check the overall accuracy of the approximator in (6) and the validity of its standard error (7) by cross validation (see Jones et al., 1998).
3. Compute the estimated marginal effects defined in (18) by carrying out the required integrations as described in Appendix B. This will usually be done for all main effects and all two-variable joint effects.
4. For each estimated marginal effect, compute the corresponding estimated corrected effect by subtracting all estimated lower-order corrected effects. This is best done recursively, correcting the main effects first, then correcting the two-variable effects, and so on.
5. Using the estimated corrected effects, compute the estimated contributions in the functional analysis of variance (14).
6. If an estimated interaction effect makes a substantial contribution to the ANOVA decomposition, the corresponding joint effect (18) is plotted against the relevant input variables as a contour or perspective plot. The standard error (19) can also be plotted against the same input variables in a separate plot.
7. Any input variable that has a large ANOVA contribution from an estimated (corrected) main effect but does not appear in any large ANOVA contributions from interaction effects has its estimated (uncorrected) main effect plotted. Approximate pointwise confidence intervals based on the standard error can also be shown.

## 6 Application: The Wonderland Model

We illustrate these methods using the Wonderland computer model outlined in Section 1. This model exemplifies the type of screening problem we have in mind, as we show that it has highly nonlinear, interactive effects that demand a flexible, data-adaptive statistical modeling strategy. The computer model has 41 input variables and we focus on one particular quasi global human development index (HDI) output variable resulting from a policy that might be called "limits to growth". The data consist of 500 model runs from a "space-filling" Latin hypercube design (see McKay et al., 1979).

The first step in the analysis is to fit the random-function model (1) and to check the accuracy of the resulting approximator. We use a simple random-function model:

$$Y(\boldsymbol{x}) = \beta_0 + Z(\boldsymbol{x}),$$

where the regression component is just a constant, $\beta_0$. The unknown parameters, $\beta_0$, $\sigma^2$ in (2) and the correlation parameters $\theta_j$ and $p_j$ for $j = 1, \ldots, 41$ in (3), are estimated by maximum likelihood. Figure 2(a) shows the actual HDI value $y(\boldsymbol{x}^{(i)})$ from run $i$ of the Wonderland model versus its leave-one-out cross-validated prediction, $\hat{Y}_{-i}(\boldsymbol{x}^{(i)})$ for $i = 1, \ldots, 500$. The subscript $-i$ indicates that the approximator (6) is built from all the data except run $i$. (The random-function correlation parameters are not re-estimated). Figure 2(a) shows fairly good accuracy of

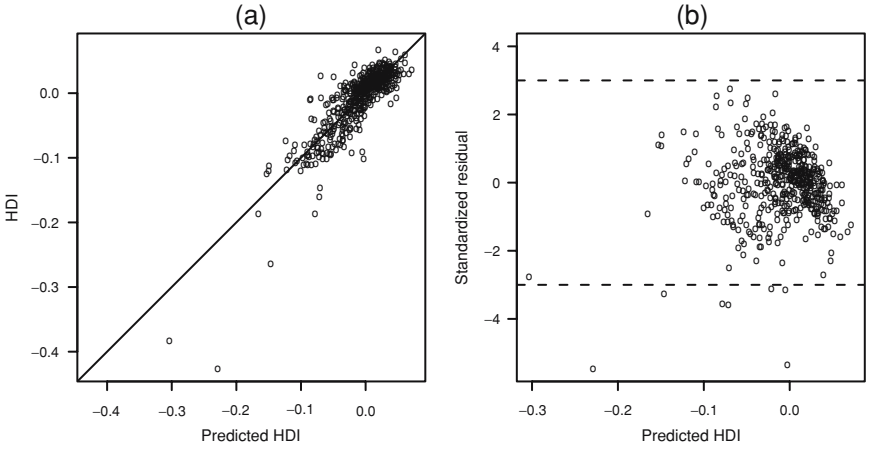



FIGURE 2. Diagnostics for the Wonderland approximating model: (a) actual human development index (HDI) values versus their cross-validation predictions; (b) standardized cross-validation residuals versus cross-validation predictions.

approximation, though with some over-prediction of the extremely low HDI values. Figure 2(b) plots the standardized cross-validated residual,

$$[y(\boldsymbol{x}^{(i)}) - \hat{Y}_{-i}(\boldsymbol{x}^{(i)})]/\mathrm{se}_{-i}[\hat{Y}(\boldsymbol{x})], \tag{20}$$

versus $\hat{Y}_{-i}(\boldsymbol{x}^{(i)})$ for $i = 1, \ldots, 500$, where the standard error $\mathrm{se}_{-i}[\hat{Y}(\boldsymbol{x})]$ is computed from (7), again without the data from run $i$. The plot shows some standardized residuals falling outside the bands at $\pm 3$, indicating that the error of approximation is sometimes a little larger in magnitude than is suggested by the standard error.

The next step is to compute the estimated marginal effects (18). Following our usual practice, this is done for all main effects and all two-variable joint effects. The required integrations over the remaining 40 or 39 variables, respectively, are computed as described in Appendix B.

Each estimated marginal effect leads to the estimate of the corresponding corrected effect in (12) or (13). This is done recursively: the estimated main effects are corrected first, followed by the two-variable interaction effects.

The functional analysis of variance in (14) is then computed from the estimated corrected effects. Here, the 41 main effects and 820 two-factor-interaction effects together account for about 89% of the total variance of the predictor. Hence, about 11% of the predictor's total variability is due to higher-order effects. Table 1 shows the estimated main effects and interaction effects that contribute at least 1% to the functional ANOVA: These 12 effects together account for about 74% of the total variation. Only six variables appear in these 12 effects; they are described in Table 2.

The ANOVA suggests that e.inov.n (economic innovation in the north) is an important input variable. Its estimated main effect in Figure 3(a) shows a strong, approximately linear trend. The estimated increase in HDI is fairly substantial:

TABLE 1. Estimated main effects and two-variable interaction effects accounting for more than 1% of the total variance of the predictor; the variable names are defined in Table 2; suffix ".n" or ".s" indicates the northern region or the southern region, respectively

| Effect | % of Total variance | Effect | % of Total variance |
|---|---|---|---|
| e.inov.n | 24.3 | v.spoll.s × v.drop.s | 2.7 |
| v.spoll.s | 13.5 | e.grth.n × e.inov.n | 1.9 |
| e.inov.s | 12.1 | v.drop.s | 1.9 |
| e.cinov.s | 5.3 | e.finit.s | 1.5 |
| v.spoll.s × v.cfsus.s | 4.6 | e.inov.n × e.inov.s | 1.4 |
| v.drop.s × v.cfsus.s | 3.7 | v.cfsus.s | 1.2 |

about 0.06 over the e.inov.n range. This was not obvious from the scatter plot in Figure 1(a); certainly any guess as to the magnitude of the increase would have been much smaller. The relationship was masked by other variables.

The estimated main effect of v.spoll.s (sustainable pollution in the south) in Figure 3(b) confirms the same nonlinearity that we could see in the scatter plot in Figure 1(b). The drop in HDI over the first twentieth of the range of v.spoll.s is substantial. Given that we sampled 500 points we would suspect roughly one-twentieth or 25 of the HDI values to be low. However, the scatter plot in Figure 1(b) shows only three low points. This hints at a highly local interaction.

The analysis of variance in Table 1 does indeed identify several estimated interaction effects involving v.spoll.s; the largest is that with v.cfsus.s (change in sustainable pollution in the south). Figure 4(a) shows the estimated joint effect of these two input variables on HDI. The surface is fairly flat for most of the area. As previously seen in the main effect plot, HDI increases rapidly with v.spoll.s when v.spoll.s is close to its lower limit. Now we see that this increase is larger for high values of v.cfsus.s (an increase from $-0.12$ to 0) than for low values of v.cfsus.s (an increase from $-0.06$ to 0). This difference appears to be substantial relative to the standard errors shown in Figure 4(b), which are roughly of order 0.01.

For comparison, we also use stepwise regression to select variables, specifically the R function step (R Development Core Team, 2005), which uses Akaike's information criterion (see Akaike, 1973). The selection from all 41 input variables

TABLE 2. Wonderland input variables that appear in the important estimated effects of Table 1. Prefix "e." or "v." indicates an economic or environmental variable, respectively

| Variable | Description |
|---|---|
| e.finit | Flatness of initial decline in economic growth |
| e.grth | Base economic growth rate |
| e.inov | Innovation rate |
| e.cinov | Effect of innovation policies (pollution taxes) on growth |
| v.spoll | Sustainable pollution |
| v.cfsus | Change in level of sustainable pollution when natural capital is cut in half |
| v.drop | Rate of drop in natural capital when pollution flows are above the sustainable level |

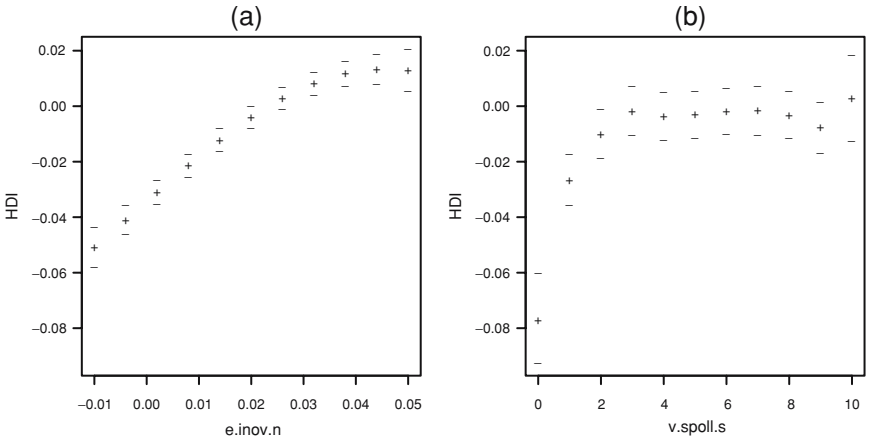



FIGURE 3. Estimated main effects on HDI in the Wonderland model: (a) estimated main effect of e.inov.n (economic innovation in the north); (b) estimated main effect of v.spoll.s (sustainable pollution in the south). The estimated effects are denoted by "+" and approximate 95% pointwise confidence limits are denoted by "−".

results in a first-order model (main effects model) with 15 variables, but e.finit.s and v.cfsus.s in Table 1 are not included. Extending the model search space to allow all second-order terms also, that is, the 41 squares and 820 bilinear interaction effects of the input variables, yields a final model with 62 terms. Again e.finit.s and v.cfsus.s do not appear. Thus, the bilinear v.spoll.s × v.cfsus.s interaction effect is not included, contrary to Table 1. (Note that a two-factor interaction effect is defined to be a more general, nonadditive effect in the random-function model.)

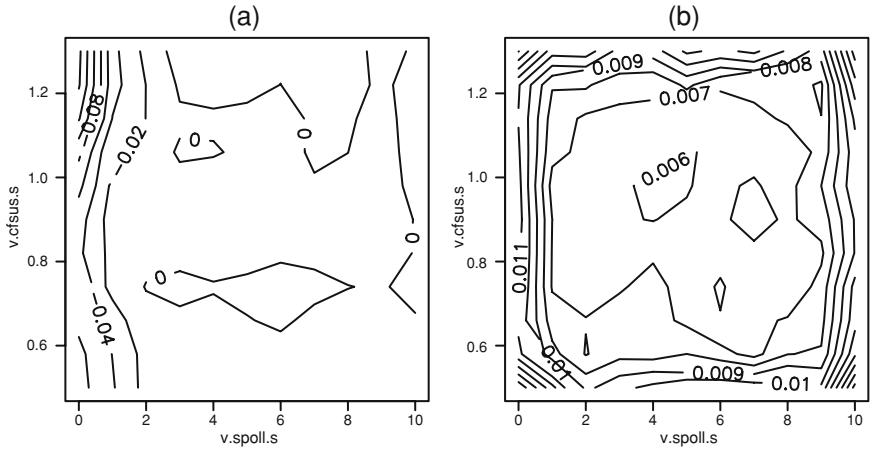



FIGURE 4. Joint effect of sustainable pollution in the south (v.spoll.s) and change in sustainable pollution in the south (v.cfsus.s) on HDI in the Wonderland model: (a) estimated effect; (b) pointwise standard error of the estimated effect.

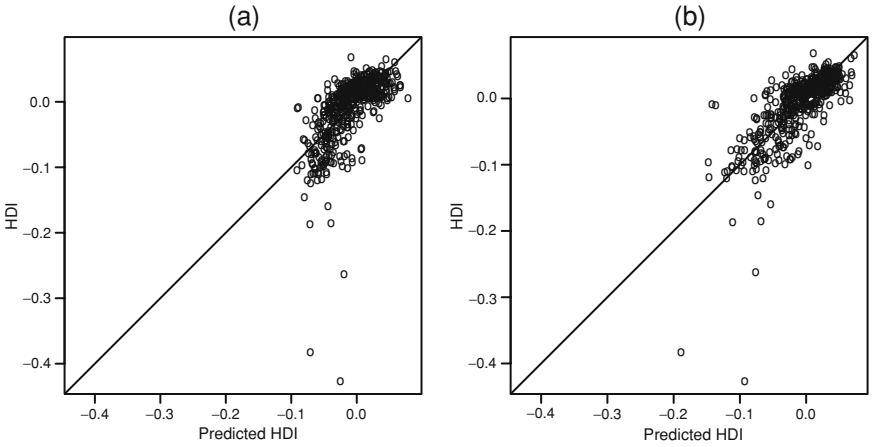



FIGURE 5. Actual human development index (HDI) values versus their cross-validation predictions from regression models: (a) selected from all first-order terms; (b) selected from all second-order terms.

The two regression models have lower prediction accuracy than the random-function model when assessed using "cross-validated root mean squared error of prediction". This quantity is simply the root mean of the squared cross-validated residuals $y(\boldsymbol{x}^{(i)}) - \hat{Y}_{-i}(\boldsymbol{x}^{(i)})$, for $i = 1, \ldots, n$ in the numerator of (20). The cross-validated root mean squared error values are 0.040 and 0.035 for the first-order and second-order regression models, respectively, compared with 0.026 for the random-function model. Figure 5 shows that both regression models are particularly poor at predicting extremely low values of HDI. The true relative importances of the effects of the input variables are not known for the Wonderland model, but it is arguable that the screening results from the random-function model are more credible because of the better prediction performance of this model.

## 7 Discussion

The Wonderland model illustrates that, at least in some applications, very complex effects involving highly nonlinear, interactive relationships, can exist. Naturally, these are difficult to model and identify. The approach that we have described starts with a random-function model that is data-adaptive to such complexities, given enough computer-model runs. Similarly, the estimated effects derived from the random-function approximator can be fairly complex if demanded by the data. To detect such subtle effects, the experimental design has to allow exploration of the input region densely, at least for a few variables at a time. "Space-filling" designs such as Latin hypercubes, used in this chapter, and orthogonal arrays (see Chapter 7) have good projective properties and are desirable in this respect (see Koehler and Owen, 1996, for a review of designs for computer experiments). The design does not have to be balanced in any sense. The ANOVA decomposition is

of the approximator (that is, the predictor from the surrogate model); it is not a traditional analysis of variance computed directly from the data.

In the Wonderland model, a pathological scenario was identified of very low values of the human development index, dependent on extreme values of two of the 41 variables. In the experiment with 500 runs, only three runs exhibited this behavior; fewer runs in the design or a less flexible approximation strategy may well have missed this feature.

In practice, one is often faced with choosing a model that is easily interpretable but may not approximate a response very well, such as a low-order polynomial regression, or with choosing a black box model, such as the random-function model in equations (1)–(3). Our approach makes this black box model interpretable in two ways: (a) the ANOVA decomposition provides a quantitative screening of the low-order effects, and (b) the important effects can be visualized. By comparison, in a low-order polynomial regression model, the relationship between input variables and an output variable is more direct. Unfortunately, as we have seen, the complexities of a computer code may be too subtle for such simple approximating models.

Throughout, we have used "plug-in" estimates of the correlation parameters in (3). These parameters are estimated by maximum likelihood but the estimates are treated as the true values thereafter. The uncertainty from estimating the correlation parameters is not propagated through to the standard errors of estimated effects. In principle, though, this is easily overcome with a Bayesian prior distribution on the correlation parameters. We could: (1) sample say 10–100 sets of values of the correlation parameters from their Bayesian posterior distribution (see, for example, Robert and Casella, 2004); (2) estimate effects using the methods in this chapter, conditional on each set of values of the correlation parameters; and (3) combine the analyses using standard probability results to compute a standard error taking account of parameter uncertainty. As the analysis in this chapter is relatively straight-forward computationally, repeating it 10–100 times would not be onerous; rather, the difficulty would be with sampling from the Bayesian posterior for applications with many input variables.

## Appendix A
## Derivation of the Best Linear Unbiased Predictor of an Effect

The best linear unbiased predictor (BLUP) of $\bar{Y}_e(\boldsymbol{x}_e)$ in (18) follows from the properties of $\bar{Z}_e(\boldsymbol{x}_e)$ in (17). Clearly, $\bar{Z}_e(\boldsymbol{x}_e)$, like $Z(\boldsymbol{x})$, has expectation zero. Its variance, however, differs from one effect to another:

$$\begin{aligned}\mathrm{Var}[\bar{Z}_e(\boldsymbol{x}_e)] &= \int\limits_{\otimes_{j\notin e}\chi_j}\int\limits_{\otimes_{j\notin e}\chi_j} \mathrm{Cov}[Z(\boldsymbol{x}_e,\boldsymbol{x}_{-e}), Z(\boldsymbol{x}_e,\tilde{\boldsymbol{x}}_{-e})]\prod_{j\notin e} w_j(x_j)w_j(\tilde{x}_j)dx_j d\tilde{x}_j \\ &= \sigma^2 \int\limits_{\otimes_{j\notin e}\chi_j}\int\limits_{\otimes_{j\notin e}\chi_j} R[(\boldsymbol{x}_e,\boldsymbol{x}_{-e}),(\boldsymbol{x}_e,\tilde{\boldsymbol{x}}_{-e})]\prod_{j\notin e} w_j(x_j)w_j(\tilde{x}_j)dx_j d\tilde{x}_j. \end{aligned} \tag{21}$$

The steps for deriving the BLUP of $\bar{Y}_e(\boldsymbol{x}_e)$ with a standard error closely follow those in Section 2 for the BLUP of $Y(\boldsymbol{x})$. Again, consider predictors that are linear in the $n$ observed output values, $\hat{\bar{Y}}_e = \boldsymbol{a}'_e(\boldsymbol{x}_e)\boldsymbol{y}$. From the random-function model (15), the mean squared error of $\hat{\bar{Y}}_e(\boldsymbol{x}_e)$ is

$$\begin{aligned}\mathrm{E}[\bar{Y}_e(\boldsymbol{x}_e) - \hat{\bar{Y}}_e(\boldsymbol{x}_e)]^2 &= \mathrm{E}[\bar{\boldsymbol{f}}'_e(\boldsymbol{x}_e)\beta + \bar{Z}_e(\boldsymbol{x}_e) - a'_e(\boldsymbol{x}_e)(\boldsymbol{F}\boldsymbol{\beta} + z)]^2 \\ &= \{[\bar{\boldsymbol{f}}'_e(\boldsymbol{x}_e) - \boldsymbol{a}'_e(\boldsymbol{x}_e)\boldsymbol{F}]\boldsymbol{\beta}\}^2 + \mathrm{Var}[\bar{Z}_e(\boldsymbol{x}_e)] \\ &\quad + \boldsymbol{a}'_e(\boldsymbol{x}_e)\mathrm{Cov}(z)\boldsymbol{a}_e(\boldsymbol{x}_e) - 2\boldsymbol{a}'_e(\boldsymbol{x}_e)\mathrm{Cov}[\bar{Z}_e(\boldsymbol{x}_e), z].\end{aligned}$$

Element $i$ of the $n \times 1$ vector $\mathrm{Cov}[\bar{Z}_e(\boldsymbol{x}_e), z]$ is computed from

$$\begin{aligned}\mathrm{Cov}[\bar{Z}_e(\boldsymbol{x}_e), Z(\boldsymbol{x}^{(i)})] &= \int_{\otimes_{j\notin e}\chi_j} \mathrm{Cov}[Z(\boldsymbol{x}_e, \boldsymbol{x}_{-e}), Z(\boldsymbol{x}_e^{(i)}, \boldsymbol{x}_{-e}^{(i)})] \prod_{j\notin e} w_j(x_j)dx_j \\ &= \sigma^2 \int_{\otimes_{j\notin e}\chi_j} R[(\boldsymbol{x}_e, \boldsymbol{x}_{-e}), (\boldsymbol{x}_e^{(i)}, \boldsymbol{x}_{-e}^{(i)})] \prod_{j\notin e} w_j(x_j)dx_j. \quad (22)\end{aligned}$$

Thus, we have to integrate out the variables not in $\boldsymbol{x}_e$ from the correlation function. We write $\sigma^2\bar{\boldsymbol{r}}_e(\boldsymbol{x}_e)$ for $\mathrm{Cov}[\bar{Z}_e(\boldsymbol{x}_e), z]$.

Again imposing a constraint to eliminate the contribution to the mean squared error from the term involving $\boldsymbol{\beta}$, the optimal choice of $\boldsymbol{a}'_e(\boldsymbol{x}_e)$ is formulated as

$$\min_{\boldsymbol{a}_e(\boldsymbol{x}_e)} \mathrm{Var}[\bar{Z}_e(\boldsymbol{x}_e)] + \sigma^2\boldsymbol{a}'_e(\boldsymbol{x}_e)\boldsymbol{R}\boldsymbol{a}_e(\boldsymbol{x}_e) - 2\sigma^2\boldsymbol{a}'_e(\boldsymbol{x}_e)\bar{\boldsymbol{r}}_e(\boldsymbol{x}_e) \quad (23)$$

subject to

$$\boldsymbol{F}\boldsymbol{a}_e(\boldsymbol{x}_e) = \bar{\boldsymbol{f}}_e(\boldsymbol{x}_e).$$

The constrained optimization problems leading to the BLUPs of $Y(\boldsymbol{x})$ and $\bar{Y}_e(\boldsymbol{x}_e)$ are very similar: $\mathrm{Var}[Z(x)]$, $\boldsymbol{r}(\boldsymbol{x})$, and $\boldsymbol{f}(\boldsymbol{x})$ in (5) have simply been replaced by $\mathrm{Var}[\bar{Z}_e(\boldsymbol{x}_e)]$, $\bar{\boldsymbol{r}}_e(\boldsymbol{x}_e)$, and $\bar{\boldsymbol{f}}_e(\boldsymbol{x}_e)$, respectively, in (23).

## Appendix B
## Computation of the Integrals Required for the Estimated Effects and the ANOVA Decomposition

To compute the BLUP (18) of a marginal effect we need the vectors $\bar{\boldsymbol{f}}_e(\boldsymbol{x}_e)$ in (16) and $\bar{\boldsymbol{r}}_e(\boldsymbol{x})$ following (22), both of which involve integration over all the variables not in $e$. For computational convenience in performing these integrations, we need two further "product-structure" conditions, in addition to (8) and (9). They relate to the properties of the random-function model, specifically the regression functions, $\boldsymbol{f}(\boldsymbol{x})$, in (1) and the correlation function, $R(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, in (2).

First, we assume each regression function is a product of functions in just one input variable; that is, element $k$ of $\boldsymbol{f}(\boldsymbol{x})$ can be written

$$f_k(\boldsymbol{x}) = \prod_{j=1}^{d} f_{kj}(x_j) \qquad (k = 1, \ldots, h). \tag{24}$$

Fortunately, the polynomial regression models commonly used are made up of functions $f_k(\boldsymbol{x})$ that are products of powers of single variables. With (24), element $k$ of $\bar{\boldsymbol{f}}_e(\boldsymbol{x}_e)$ in (16) is

$$\int_{\otimes_{j\notin e}\chi_j} f_k(\boldsymbol{x}_e, \boldsymbol{x}_{-e}) \prod_{j\notin e} w_j(x_j)dx_j = \prod_{j\notin e} f_{kj}(x_j) \int_{\otimes_{j\notin e}\chi_j} \prod_{j\notin e} f_{kj}(x_j)w_j(x_j)dx_j.$$

The integral on the right-hand side of this equation is clearly a product of one-dimensional integrals,

$$\int_{\chi_j} f_{kj}(x_j)w_j(x_j)dx_j,$$

which can be evaluated using simple techniques such as Simpson's rule.

Second, we assume similarly that the correlation function is a product of one-dimensional correlation functions; that is,

$$R(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \prod_{j=1}^{d} R_j(x_j, \tilde{x}_j). \tag{25}$$

The power-exponential correlation function (3), for example, is of this product form. To compute $\bar{\boldsymbol{r}}_e(\boldsymbol{x})$, the integral on the right-hand side of (22) is evaluated as

$$\prod_{j\in e} R_j(x_j, x_j^{(i)}) \int_{\otimes_{j\notin e}\chi_j} \prod_{j\notin e} R_j(x_j, x_j^{(i)})w_j(x_j)dx_j,$$

and the integral involved is a product of one-dimensional integrals,

$$\int_{\chi_j} R_j(x_j, x_j^{(i)})w_j(x_j)dx_j.$$

For the standard error (19), we also need $\mathrm{Var}[\bar{Z}_e(\boldsymbol{x}_e)]$ in (21). With condition (25), the double integral on the right-hand side of (21) is computed as

$$\prod_{j\in e} R_j(x_j, x_j) \prod_{j\notin e} \int_{\chi_j} \int_{\chi_j} R_j(x_j, \tilde{x}_j)w_j(x_j)w_j(\tilde{x}_j)dx_j d\tilde{x}_j.$$

Thus, two-dimensional numerical quadrature is sufficient. Further simplification follows by noting that the correlation function should satisfy $R_j(x_j, x_j) = 1$ when modeling a continuous function.

To visualize the estimated effect $\hat{\bar{Y}}_e(\boldsymbol{x}_e)$ and its standard error, $\mathrm{se}[\hat{\bar{Y}}_e(\boldsymbol{x}_e)]$, these quantities are computed for a grid of values of $\boldsymbol{x}_e$. The required one- and

two-dimensional integrals depend only on the variables not in $\boldsymbol{x}_e$ and need be computed only once for all grid points.

From the estimated marginal effects, it is straightforward to compute estimates of the corrected main effects (12), the two-variable interaction effects (13), and so on. The ANOVA contributions on the right-hand side of (14) for these low-order effects involve correspondingly low-dimension integrals.

*Acknowledgments.* We are indebted to Joe Hendrickson for supplying the data for the Wonderland model and answering our numerous questions about the application. Schonlau's research was supported by core funding from the RAND corporation. Welch's research was supported by the Natural Sciences and Engineering Research Council of Canada.

## *References*

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281. Akademia Kiadó, Budapest.

Aslett, R., Buck, R.J., Duvall, S.G., Sacks, J., and Welch, W.J. (1998). Circuit optimization via sequential computer experiments: Design of an output buffer. *Journal of the Royal Statistical Society, C*, **47**, 31–48.

Bernardo, M.C., Buck, R., Liu, L., Nazaret, W.A., Sacks, J., and Welch, W.J. (1992). Integrated circuit design optimization using a sequential strategy. *IEEE Transactions on Computer-Aided Design*, **11**, 361–372.

Bettonvil, B. and Kleijnen, J.P.C. (1996). Searching for important factors in simulation models with many factors: Sequential bifurcation. *European Journal of Operational Research*, **96**, 180–194.

Chapman, W.L., Welch, W.J., Bowman, K.P., Sacks, J., and Walsh, J.E. (1994). Arctic sea ice variability: Model sensitivities and a multidecadal simulation. *Journal of Geophysical Research C*, **99**, 919–935.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, **86**, 953–963.

Gough, W.A. and Welch, W.J. (1994). Parameter space exploration of an ocean general circulation model using an isopycnal mixing parameterization. *Journal of Marine Research*, **52**, 773–796.

Gu, C. and Wahba, G. (1993). Smoothing spline ANOVA with componentwise Bayesian "confidence intervals". *Journal of Computational and Graphical Statistics*, **2**, 97–117.

Herbert, R.D. and Leeves, G.D. (1998). Troubles in Wonderland. *Complexity International*, **6**. http://www.complexity.org.au/ci/vol06/herbert/herbert.html

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, **19**, 293–325.

Iman, R.L. and Conover, W.J. (1980). Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. *Communications in Statistics A—Theory and Methods*, **9**, 1749–1842.

Jones, D.R., Schonlau, M., and Welch, W.J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, **13**, 455–492.

Koehler, J.R. and Owen, A.B. (1996). Computer experiments. In *Handbook of Statistics*, Volume 13. Editors: S. Ghosh and C.R. Rao. Elsevier, Amsterdam.

Lempert, R.J., Popper, S.W., and Bankes, S.C. (2003). *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-term Policy Analysis*. RAND, Santa Monica, CA. http://www.rand.org/publications/MR/MR1626

McKay, M.D., Conover, W.J., and Beckman, R.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239–245.

Morris, M.D. (1991). Factorial sampling plans for preliminary computer experiments. *Technometrics*, **33**, 161–174.

Mrawira, D., Welch, W.J., Schonlau, M., and Haas, R. (1999). Sensitivity analysis of computer models: World Bank HDM-III model. *Journal of Transportation Engineering*, **125**, 421–428.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. http://www.R-project.org

Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, second edition. Springer, New York.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989). Design and analysis of computer experiments (with discussion). *Statistical Science*, **4**, 409–435.

Santner, T.J., Williams, B.J., and Notz, W.I. (2003). *The Design and Analysis of Computer Experiments*. Springer Verlag, New York.

Schonlau, M. (1997). Computer experiments and global optmization. PhD thesis, University of Waterloo, Waterloo, Ontario.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., and Morris, M.D. (1992). Screening, predicting, and computer experiments. *Technometrics*, **34**, 15–25.

# Index


$2^{f-q}$ design, 16
$2^f$ design, 3

A-efficiency, 173
aberration, 172
active control, 144
active factors, 29, 172, 194
active learning, 72
adaptive, 268, 274
adaptive variance estimator, 283
ADME, 70
aggregation, 295
Akaike's information criterion, 320
aliases, 8, 9, 265
aliasing, 8
all subsets regression, 238
all subsets selection, 180
analysis of variance (ANOVA), 4, 14, 311, 314, 316–319, 320, 322, 326
antagonism, 51
approximation error, 293
approximator, 313
assay calibration, 69
Average Entropy, 127

balanced design, 217
base level, 289
base value, 289, 294
basic design, 7, 8
basis functions, 288
bayes factor, 182
Bayesian analysis, 242
Bayesian modeling, 170
Bergman–Hynén method, 33, 37, 41, 44
Bergman–Hynén statistic, 41–43
best linear unbiased predictor (BLUP), 313, 323
biased estimation, 181
bifurcation, 298
bioavailability, 106
bioisostere, 80
biological assay (biochemical test), 76
biology, 70
black box, 288
blocking, 297, 142
blood-brain barrier, 112
blood-glucose experiment, 235
BLUP, 314, 317, 324
Box–Meyer method, 31, 32, 41, 44
Box–Meyer statistic, 41–43
breakdown point, 272, 275
Brenneman–Nair method, 39, 44
burn-in, 248

candidate design, 210
categorical factor, 18
cell-based design, 73
center points, 6
chemical diversity, 69
Cheminformatics, 76
chemistry, 70
classification, 49, 50, 52–54, 56, 62
closed testing, 146, 281
coded level, 294, 305
coded units, 3
columnwise-pairwise algorithms, 226
combination therapies, 63
combined arrays, 27
common pseudorandom numbers, 297, 298, 302

completely randomized design, 185
composite sampling, 49
compound noise factors, 27
compound orthogonal arrays, 28
computer experiments, 72, 309, 312
computer models, 308, 318
confirmation experiments, 11
conjugate prior, 242
contour plot, 16
contrasts, 3
control factors, 21
controllable factors, 305
corrected effects, 315, 316, 318, 319
correlation function, 312, 325
coverage design, 84
cross validation, 318
cross-product array, 27
cross-validated residuals, 319, 322
cross-validation, 99, 182
crossed array, 27
curvature in the response function, 5, 18
cycle design, 179

D-efficiency, 173
D-optimal design, 83
data cleaning, 89
defining relation, 8, 159
defining word, 159, 161
descriptor (property, covariate, predictor), 71
design catalogue, 215
design factors, 21, 25
design for model selection, 263
design generator, 7
design matrix, 3, 209
design projection, 11
design region, 288
design space, 211
diagnostic, 140
discrepancy, 172
dispersion effect, 21, 25, 26, 28, 31, 36, 38–44
dispersion model, 32, 37, 39, 42
distance metric, 79
diversity, 72
Dorfman retesting, 53
drug discovery, 69
drugability, 105

EC-optimal designs, 209, 217
EC50, 73
effect heredity, 238, 244
effect hierarchy, 244
effect screening, 203, 236
effect sparsity, 10, 156, 191, 244, 270, 288
efficiency, 294
environmental factor, 294, 305
enzyme, 70
$E(s^2)$ criterion, 172
estimation, 49, 50, 53, 54
estimation capacity, 208
euclidean distance, 81
exact confidence interval, 273, 276
exchange algorithm, 178
expected number of active effects, 258
expected prediction difference, 214, 222
experimental domain, 288
exploratory data analysis, 170
expression levels, 139
expression profile, 118

factor cancellation, 196
factor effect estimates, 4
factor group, 293
factor level, 293
factor screening, 1, 2, 9, 16, 29, 191
factor sparsity, 188
factorial designs, 3
false negative, 72
false negative predictive value, 59, 60
false positive, 72
false-positive predictive value, 59, 60
Familywise Error Rate (FWER), 144, 152
first-order effect, 293
fold change, 120
foldover, 163
foldover design, 12, 228, 300
follow-up experiments, 11
food and drug administration (FDA), 108
forward selection, 181
fractional factorial designs, 179
frequentist inference, 185
full aliasing, 265
full estimation capacity design, 208
functional, 209
functional analysis of variance (ANOVA), 311, 316, 318, 319

Gamma distribution, 120
Gaussian random-function, 311
gene, 139
gene expression, 115
gene expression levels, 150
generalized Latin squares, 150
generalized least squares, 314
generalized linear model, 25, 264

genetic algorithm, 178
Gibbs-sampling, 182, 247
group screening, 183, 192

Hadamard matrices, 160, 173
half-normal probability plot of effects, 5, 13, 29
Harvey Method, 34, 35, 40, 41, 43, 44
hidden projection, 162
high-throughput screening (HTS), 62, 63, 69
hit compounds, 49
hit rate, 75
hits, 73
HIV, 54, 56, 57, 62, 63
housekeeping genes, 143
hyperparameters, 257

IC50, 73
imputation, 90
inactive factor, 173
incomplete block design, 173
industrial screening experiments, 2
information capacity, 212
input variables, 308, 312
interaction effects, 316, 318, 319, 321, 326
interactions, 293, 300, 303
interaction screening, 203
interchange algorithm, 178
interpretation fractional factorials, 11–13, 17
intrinsic Bayes factor, 182
isomorphic designs, 215
iterated fractional factorial designs, 290
iterative analysis, 271

joint effects, 315, 316, 318–21

Kriging, 288

Latin hypercube, 166, 309, 310, 318, 322
Latin hypercube sampling, 305
lead compounds, 62
learning curve, 97
least squares, 4
Lenth method, 5, 29, 42
level-$\alpha$ test, 273
linear model, 293
linear predictor, 313
location effects, 28, 31, 37, 38, 41, 42, 44
location models, 29, 31, 32, 35–38, 40–44
log-linear model, 32, 34, 36, 37, 39, 42, 43
lognormal distribution, 120
lower bounds, 173

machine learning, 70
Mahalanobis distance, 80
main effects, 2, 3, 293, 294, 315, 316, 318, 319, 321, 326
marginal effects, 315, 316, 318, 319, 326
marginal likelihood, 246
masking, 309
maximum likelihood, 312, 317, 318
maximum likelihood estimator, 35, 36
maximum prediction difference, 214
McGrath–Lin parametric method, 37
McGrath–Lin statistic, 41–44
McGrath–Lin's nonparametric method, 38
MCMC, 247, 262
mean orthogonal design, 217
mean squared error, 313, 322, 324
medium-throughput screening (MTS), 73
metabolism, 69
metamodels, 288
microarray experiment, 118, 139
microarrays, 115
minimum aberration, 10, 172
mirror observation, 300
mixed linear models, 39
mixed resolution, 28
mixed-level design, 18
mixture distributions, 243
model assessment, 101
model discrimination, 209, 223
model matrix, 209
model selection, 89, 185
model space, 95, 211
model-robust factorial design, 212
model-robust orthogonal design, 217
model-robust parameter design, 221
molecular diversity, 70
molecule (compound), 69
most powerful test, 273
Multi-Objective Optimization, 109
multicriteria decision making, 72
multiple regression, 173
multiple grouping, 200
multiple linear regression (MLR), 92
multiple outputs, 303, 304
multiple stage screening, 196

negative control, 144
nested halving, 54
neural net, 288
noise factors, 21, 22, 25, 28, 305

non-regular design, 219, 237
nongeometric design, 17
nonlinear effects, 309
nonlinear model, 293
nonorthogonal, 283
nonregular designs, 160, 264
normal probability plot of effects, 5
normal probability plot of residuals, 15
null case, 273
null distribution, 274

one-factor-at-a-time, 290
optimization, 1, 305
orthogonal, 176, 197
orthogonal arrays, 27, 157
orthogonal designs, 210, 215, 269
orthogonal pooling, 50
output variables, 308, 312
overfitting, 93

Paley design, 163
partial aliasing, 17, 160
partial foldover, 12
partition testing, 146, 147
penalized least squares, 181
permutation test, 143
Plackett–Burman design, 17, 160, 174
polynomial models, 293
polynomial regression, 288
pooling, 72
positive predictive value, 105
positive discovery rate, 105
posterior distribution, 187, 246
posterior probability, 182, 248
potency, 70
power-exponential correlation function, 312, 325
pre-experimental planning, 2
predictive distribution, 254
predictive Modeling, 75
principal fraction, 8
prior distributions, 182, 242
prior hyperparameters, 252
prior knowledge, 294
prior probabilities, 172
process characterization, 1
product characterization, 1
prognostic, 140
projection designs, 156
projection properties, 11, 162
projectivity, 157
property space, 71
pseudo standard error, 274
pseudorandom numbers, 290, 293, 302

quadratic effect, 305
quadrature, 325
qualitative factor, 3, 18

random balance, 169
random effects, 39
random forests, 93
random function, 312, 317
randomization, 142, 185
rare events, 72
receptor, 70
recursive partitioning (tree model), 95
regression, 180
regression coefficients, 4, 172
regular fractional factorial design, 8, 158, 219
replication, 290, 142
reproducibility, 123
resampling, 181
resolution, 9, 169, 159
resolution III, 9, 12
resolution IV, 9, 12, 311
resolution V, 9
resolving power, 213
response surface, 288
response surface, 1, 40
restricted CP algorithm, 226
restricted maximum likelihood, 39, 42
reuse of runs, 197
ridge regression, 181
robust design, 21–23, 25, 28, 39
robust methods, 262, 272, 275, 305

sample molecular profile, 118
saturated design, 9, 183, 268, 270
scaled level, 294
score statistic, 36
screening, 287
screening experiments, 29, 42, 44
search designs, 164
second-order model, 6
secondary screen, 73
selection bias, 185
selectivity, 71
semi-foldover, 12
sensitivity, 52, 59, 140
sensitivity analysis, 310, 311
separate location models, 33
sequential bifurcation, 287, 296, 301, 310
sequential experimentation, 2, 289
shrinkage, 255
signal-to-noise ratio, 26
similarity, 79

simple linear regression, 172
simple pooling, 50
simulation, 70, 289
simulation software, 302
single-step tests, 280, 281
size-$\alpha$ test, 273, 276
space filling designs, 83, 165
sparsity of effects principle, 10, 288
specificity, 52, 59, 90, 140
spline, 288, 311
standard errors, 314, 317–21
standardization, 293
steady state, 303
steepest ascent, 2
step-down tests, 284, 281
step-up tests, 284
step-wise screening, 200
stepdown testing, 146
stepwise regression, 17, 237, 320
stepwise selection, 181
stochastic search variable selection, 182
Stochastic Ordering Lemma, 277, 283
strength, 157
strong control, 145, 276
strong heredity, 300
structure activity relationship (SAR, QSAR), 51, 71
studentized residuals, 15
subset pivotality, 149
subspace angle, 214, 222
supersaturated designs, 17, 169, 283, 290, 310
supply chain, 290, 302
support vector machines, 93
synergism, 51, 63

Taguchi, 21, 22
Tanimoto coefficient, 79
target, 69
test for curvature, 6
test set, 90
toxicity, 69
training set, 80
transmitted variation, 24
tree, 63
two-factor interactions, 3, 5
two-stage screening, 194

uniform designs, 165, 172
unreplicated factorial design, 5

validation, 302, 305
variable selection, 79
variance heterogeneity, 297, 298
variance reduction, 297
virtual screen, 70
visualization, 311, 317

Wang Method, 36, 44
weight function, 315
Wonderland model, 309, 318, 322, 323