Use R!

Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R

Order-Restricted Analysis of Microarray Data

Springer

Use R!

Series Editors:
Robert Gentleman Kurt Hornik Giovanni Parmigiani

For further volumes:
http://www.springer.com/series/6991

Dan Lin • Ziv Shkedy • Daniel Yekutieli
Dhammika Amaratunga • Luc Bijnens
Editors

Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R

Order-Restricted Analysis of Microarray Data

Springer

Editors
Dan Lin
Veterinary Medicine Research
& Development, Pfizer Animal Health
Zaventem, Belgium

Ziv Shkedy
Interuniversity Institute for Biostatistics
and statistical Bioinformatics (I-BioStat)
Center for Statistics (CenStat)
Hasselt University, Diepenbeek, Belgium

Daniel Yekutieli
Department of Statistics and Operations
Research, School of Mathematical Sciences,
Tel-Aviv University, Tel-Aviv, Israel

Dhammika Amaratunga
Biostatistics and Programming
Janssen Pharmaceutical Companies
of Johnson & Johnson
Raritan, NJ USA

Luc Bijnens
Biostatistics and Programming
Janssen Pharmaceutical Companies
of Johnson & Johnson
Beerse, Belgium

R programs and datasets: http:/www.ibiostat.be/software; R packages on http://www.r-project.org; http://cran.r-project.org; http://www.bioconductor.org and/or via springer.com/statistics/book/978-3-642-24006-5

ISBN 978-3-642-24006-5 ISBN 978-3-642-24007-2 (eBook)
DOI 10.1007/978-3-642-24007-2
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012945536

Preface

Bioinformatics and statistical bioinformatics have developed rapidly over the last 15 years. In particular, the development of microarray technology introduced the challenge of the analysis of massive datasets and the need to consider inference when thousands of genes are tested.

Microarray experiments are slowly becoming an integrated part of the pharmaceutical research and development (R&D) process. Microarray experiments offer the ability to measure, at the same time, the RNA derived from entire genomes. Microarrays specific for animal species can be used to process the information coming from in vivo animal experiments. Microarrays specific for humans can be used to test biological material coming from biopsies, blood samples, or cell line cultures. The functional genomic information coming from microarray experiments can be used both at the target identification and the target validation level of the early drug discovery process. Moreover, functional genomics can be used at many later stages of pharmaceutical research and development to screen therapeutic effects and unwanted side effects in many R&D programs.

This book is about a specific setting in which gene expression is measured at different dose levels of a drug. The main goal of the analysis of dose-response microarray experiments is to detect trends in gene expression caused by increasing doses of compound. Therefore, this book is focused on estimation, inference, and clustering under order restrictions of dose-response microarray data. The aim of these microarray experiments is to get insight into the mechanism of action and the safety profile of a drug using functional genomic data to identify pathways that are affected by the compound at hand. In this context, gene expression experiments have become important either before or parallel to the clinical testing programs.

In this book, we present a toolbox for the analysis of dose-response microarray experiments. The toolbox consists of different statistical methods for the analysis and different R packages which were developed for the analysis. The web site accompanying this book contains all the R programs and datasets used to produce the output presented in the book. It can be reached through the web site of Hasselt University:

http://www.ibiostat.be/software/IsoGeneGUI/index.html

R packages can be downloaded from either the R Project web site, R-Forge, CRAN, or the Bioconductor web site: http://www.r-project.org/, http://cran.r-project.org/, http://r-forge.r-project.org/, and http://www.bioconductor.org/, respectively.

In the first part of the book, we introduce the main concepts of estimation and inference under order constraints and dose-response modeling. In the second part of the book, we focus on the analysis of dose-response microarray experiments and address issues such as multiplicity adjustment, selective inference, single and multiple contrast tests, order-restricted clustering, pathway analysis, hierarchical Bayesian models, and model-based approach.

Bioinformatics and statistical bioinformatics are multidisciplinary areas. The materials presented in this book have been developed over the last few years by a group of biologists, biostatisticians, mathematical statisticians, and computer scientists from both academia and pharmaceutical industry. Most of the coauthors of this book are part of the CHIPS (common hour involving practical statistics) network. This network was initiated about 10 years ago as biweekly workshops to give statisticians, bioinformaticians, and life scientists the opportunity to integrate expertise on the design and analysis of microarray experiments. The network has a strong culture of sharing knowledge and expertise among the different professions.

Last, but not least, we would like to thank all our collaborators, without their work this book could never have been published: Marc Aerts, Frank Bretz, Tomasz Burzykowski, Djork-Arn Clevet, An De Bondt, Gemechis D. Dijra, Filip De Ridder, Hinrich. W.H. Göhlmann, Philippe Haldermans, Sepp Hochreiter, Ludwig Hothorn, Adetayo Kasim, Bernet Kato, Martin Otava, Setia Pramana, Pieter Peeters, Tim Perrera, Jose Pinheiro, Nandini Raghavan, Roel Straetemans, Willem Talloen, Suzy Van Sanden, and Tobias Verbeke. We would like to thank Niels Thomas and Alice Blanck from Springer, Heidelberg, and the head of the production team Ms. Ranjani Shanmugaraj for all their help and support during the preparation of this book.

Zaventem, Belgium — Dan Lin
Diepenbeek, Belgium — Ziv Shkedy
Tel-Aviv, Israel — Daniel Yekutieli
Raritan, USA — Dhammika Amaratunga
Beerse, Belgium — Luc Bijnens
July 2012

Contents

Contributors

Marc Aerts Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics (CenStat), Hasselt University, Diepenbeek, Belgium

Dhammika Amaratunga Biostatistics and Programming Jansenn Pharmaceutical Companies of Johnson & Johnson Raritan, NJ USA

Luc Bijnens Biostatistics and Programming Jansenn Pharmaceutical Companies of Johnson & Johnson Beerse, Belgium

Frank Bretz Integrated Information Sciences, Novartis Pharma AG, Novartis Campus, Basel, Switzerland

Tomasz Burzykowski Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics (CenStat), Hasselt University, Diepenbeek, Belgium

Djork-Arn Clevert Institute of Bioinformatics, Johannes Kepler University, Linz, Austria

An De Bondt Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium

Filip De Ridder Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium

Gemechis D. Djira Department of Mathematics and Statistics, South Dakota State University, Brookings, SD, USA

Hinrich W.H. Göhlmann Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium

Philippe Haldermans Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics (CenStat), Hasselt University, Diepenbeek, Belgium

Sepp Hochreiter Institute of Bioinformatics, Johannes Kepler University, Linz, Austria

Ludwig A. Hothorn Institute of Biostatistics, Leibniz University Hannover, Hannover, Germany

Adetayo Kasim Wolfson Research Institute, Durham University, Stockton, UK

Bernet Kato Respiratory Epidemiology and Public Health, Imperial College London, London, UK

Dan Lin Veterinary Medicine Research and Development, Pfizer Animal Health, Zaventem, Belgium

Martin Otava Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics (CenStat), Hasselt University, Diepenbeek, Belgium

Pieter Peeters Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium

Tim Perrera Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium

Jose Pinheiro Janssen Pharmaceutical Companies of Johnson & Johnson, Titusville, NJ, USA

Setia Pramana Department of Medical Epidemiology and Biostatistics Karolinska Institutet, Stockholm, Sweden

Nandini Raghavan Janssen Pharmaceutical Companies of Johnson & Johnson, Raritan, NJ, USA

Ziv Shkedy Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics (CenStat), Hasselt University, Diepenbeek, Belgium

Roel Straetemans Ablynx NV, Zwijnaarde, Belgium

Willem Talloen Biostatistics and Programming Jansenn Pharmaceutical Companies of Johnson & Johnson Beerse, Belgium

Ilse Van Den Wyngaert Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium

Suzy Van Sanden Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium

Tobias Verbeke OpenAnalytics BVBA, Heist-op-den-Berg, Belgium

Daniel Yekutieli Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel

R Packages

`DoseFinding`
`DoseFinding` is a CRAN package for the design and the analysis of dose-finding experiments. It provides functions for multiple contrast tests, nonlinear dose-response modeling, calculating optimal designs, and an implementation of the MCPMod methodology discussed in Pinheiro et al. (2006).

`fdrame`
`FDR-AME` is a Bioconductor package that computes FDR adjustments for p-values generated in multiple hypotheses testing of gene expression data obtained by a microarray experiment. It applies both theoretical-distribution-based and resampling-based multiple testing procedures and presents as output of the adjusted p-values and p-value plots, as described in Reiner et al. (2003).

`IsoGene`
A CRAN R package for testing monotone trends in dose-response microarray experiments. The package provides several testing procedures discussed in Lin et al. (2007). Inference is based on either the asymptotic distribution of the likelihood ratio test statistic or resampling-based inference for the t-type test statistics. Adjustment for multiplicity is based on either the BH-FDR procedure or SAM.

`IsoGeneGUI`
A graphical user interface for the `IsoGene` package that does not require an extensive knowledge of R. The package performs all the statistical tests implemented in the `IsoGene` and provides several default and user-defined graphical and numerical output. The capacity of the package is discussed in Pramana et al. (2010a,b).

`limma`
The Bioconductor `Limma` package (Smyth, 2004) fits a hybrid frequentist/eBayes linear model for the expression levels of the genes in the array. The package can be used to analyze gene expression data obtained from several microarray platforms such as two-color cDNA (including normalization function for data preprocessing) and Affymetrix.

`MLP`
A Bioconductor R package for analysis of data from a microarray experiment to determine significant sets of genes that are functionally related or in a certain biological pathway. The package performs gene set analysis using the MLP approach described in Raghavan et al. (2006). Genes are mapped into gene sets or pathways by utilizing gene annotation databases such as the Gene Ontology, KEGG, etc. The *p*-values corresponding to genes in a gene set are used to define a gene set statistic. Gene set significance is determined using a permutation procedure based on randomly reassigning *p*-values to genes.

`mratios`
The `mratios` package provides simultaneous inferences for ratios of linear combinations of coefficients in the general linear model. It includes several multiple comparison procedures as applied to ratio parameters, parallel-line and slope-ratio assays, and tests for noninferiority and superiority based on relative thresholds (Dilba et al. 2007).

`multtest`
The Bioconductor package `multtest` uses resampling-based multiple testing procedures for controlling the family-wise error rate (FWER), generalized family-wise error rate (gFWER), and false discovery rate (FDR). Single-step and stepwise methods are implemented. The results are reported in terms of adjusted *p*-values, confidence regions, and test statistic cutoffs. The procedures are directly applicable to identifying differentially expressed genes in DNA microarray experiments. The package is discussed by Dudoit and van der Laan (2008), *Multiple testing procedures with applications to genomics*.

`multcomp`
A CRAN R package for simultaneous tests and confidence intervals for general linear hypotheses in parametric models, including linear, generalized linear, linear mixed effects, and survival models. The package capacity is described in Bretz et al. (2010), *Multiple comparisons using R*.

`nlme`
A CRAN R package for fitting linear mixed models, nonlinear mixed effects models and generelaized linear models. The function `gnls()` can be used to fit nonlinear models. An elaborate discussion about the methodology is given by Pinheiro and Bates (2000), *Mixed effects models in S and S plus*.

`ORCME`
A CRAN R package for simple order-restricted clustering of dose-response microarray data. The `ORCME` package finds clusters of genes with co-regulated dose-response relationship. This package implements a variation of biclustering algorithms of Cheng and Church (2000).

`ORIClust`
An R package for order-restricted clustering of dose-response microarray data. The clustering algorithm implemented in `ORIClust`, ORICC, is a model

selection-based algorithm in which an order-restricted information criterion, proposed by Liu et al. (2009a, 2009b), is used for clustering.

`ORIOGEN` 3.0
`ORIOGEN` 3.0 (Peddada et al. 2005) is not an R package but a java-based interface which can be used to test the null hypothesis of no dose effect against order-restricted alternatives. The methodology implemented in `ROIOGEN` is discussed in Peddada et al. (2003, 2005). The package can be downloaded freely from http://dir.niehs.gov/dirbb/oriogen/index.cfm.

`pava`, `isoreg`, and `monoreg`
`pava` is an R function (from the `iso` package) which calculates a weighted isotonic regression, and `monoreg` is a CRAN R package which fits a nonparametric monotone regression functions. Both functions fit weighted isotonic regression models. Both `pava` and `monoreg` can estimate increasing (isotonic) and decreasing (antitonic) models. An elaborate discussion about the implementation of the Pool-Adjacent-Violators Algorithm (PAVA) in R can be found in De Leeuw et al. (2009).

`R2WinBUGS`
The R2WinBUGS package (Sturtz et al. 2005) consists of a set of R functions which can be used in order to call WinBUGS from R. The package automatically writes the data and scripts in a format readable by WinBUGS.

`samr`
`samr` is a CRAN package that performs the significance analysis of microarray (SAM; Tusher et al. 2001). The SAM is a statistical method for finding significant genes in a set of microarray experiments. The package allows for several types of response variables such as a two-class or multiclass grouping, a quantitative variable, or a censored survival time. The SAM is a resampling-based procedure in determining the significance of the tests and estimating the false positive rate.

References

Bretz, F., Hothorn, T., & Westfall P. (2010). *Multiple comparisons using R*. Boca Raton: CRC.

Cheng, Y., & Church, G.M. (2000). Biclustering of expression data. *Proceedings of the Conference on Intelligent Systems for Molecular Biology, 55*, 93–104.

De Leeuw, J., Hornik, K., & Mair, P. (2009). Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods. *Journal of Statistical Software, 32*(5), 1–24.

Dilba, D., Schaarschmidt, F., & Hothorn, L.A. (2007). Inferences for ratios of normal means. *R News 7, 1*, 20–23.

Dudoit, S., & van der Laan, M.J. (2008). *Multiple testing procedures with applications to genomics*. New York: Springer.

Lin, D., Shkedy, Z., Yekutieli, D., Burzykowki, T., Göhlmann, H.W.H., De Bondt, A., et al. (2007). Testing for trend in dose-response microarray experiments: Comparison of several testing procedures, multiplicity, and resampling-based inference. *Statistical Application in Genetics and Molecular Biology, 6*(1). Article 26.

Liu, T., Lin, N., Shi, N., & Zhang, B. (2009a). *Order-restricted information criterion-based clustering algorithm. Reference manual*. http://cran.r-project.org/web/packages/ORIClust/.

Liu, T., Lin, N., Shi, N., & Zhang, B. (2009b). Information criterion-based clustering with order restricted candidate profiles in short time-course microarray experiments. *BMC Bioinformatics 10*, 146.

Peddada, S., Lobenhofer, E.K., Li, L., Afshari, C.A., Weinberg, C.R., & Umbach, D.M. (2003). Gene selection and clustering for time-course and dose-response microarray experimants using order-restricted inference. *Bioinformatics, 19*(7), 834–841.

Peddada, S., Harris, S., & Harvey E. (2005). ORIOGEN: Order restricted inference for ordered gene expression data. *Bioinformatics, 21*(20), 3933–3934.

Pinheiro, J., & Bates, D. (2000). Mixed-effects models in S and S-PLUS. Springer, New York

Pinheiro, J.C., Bretz, F., & Branson, M. (2006). Analysis of dose-response studies—Modeling approaches. In N. Ting (Ed.), *Dose finding in drug development* (pp. 146–171). New York: Springer.

Pramana, S. Lin, D., Haldermans, P., Shkedy, Z., Verbeke, T., Göhlmann, H., et al. (2010a). IsoGene: an R package for analyzing dose-response studies in microarray experiments. *The R Journal, 2*(1), 5–12.

Pramana, S., Lin, D., & Shkedy, Z. (2010b). IsoGeneGUI package vignette. *Bioconductor*. http://www.bioconductor.org.

Raghavan, N., Amaratunga, D., Cabrera, J., Nie, A., Qin, J., & McMillian, M. (2006). On Methods for gene function scoring as a means of facilitating the interpretation of microarray results. *Journal of Computational Biology, 13*(3), 798–809.

Reiner, A., Yekutieli, D., & Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics, 19*(3), 368–375.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology, 3*. Article 3. http://www.bepress.com/sagmb/vol3/iss1/art

Sturtz, S., Ligges, U., & Gelman, A. (2005) R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software, 12*(3), 1–16.

Tusher, V.G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrys applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences, 98*, 5116–5121.

Acknowledgments

The work of Ludwig A. Hothorn was supported by the German Science Foundation fund HO1687/9-1 and the EC FP7 ESNATS project no. 201619.

Daniel Yekutieli gratefully acknowledged the support from the Wharton School, University of Pennsylvania, for his appointment to the Harry W. Reynolds Visiting International Professorship.

Dan Lin, Ziv Shkedy, Marc Aerts, Tomasz Burzykowski, Philippe Haldermans, Martin Otava, and Setia Pramana gratefully acknowledged the support from the Belgian IUAP/PAI network "Statistical techniques and modeling for complex substantive questions with complex data."

Chapter 1
Introduction

Dan Lin, Willem Talloen, Luc Bijnens, Hinrich W.H. Göhlmann, Dhammika Amaratunga, and Roel Straetemans

1.1 Introduction

The development of new and innovative treatments for unmet medical (Barlow et al. 1972) needs is the major challenge in biomedical research. Unfortunately, for the past decade, there has been a steady decline in the number of new therapies reaching the market, despite of the increased investments in pharmaceutical R&D (FDA 2004). One of the most critical steps in a drug discovery program is target identification and validation (Sams-Dodd 2005). Good drugs are potent and specific, that is, they must have strong effects on a specific biological pathway and minimal effects on all other pathways (Marton et al. 1998). Confirmation that a compound inhibits the intended target (drug target validation) and the identification of undesirable secondary effects are among the main challenges in developing new drugs. This is the reason why dose-response experiments are pivotal in drug discovery programs. Dose-response experiments help us to understand how the drug works and to explore whether it has the desired properties of a potential novel therapy. A compound will only move further in clinical testing when it has a side

D. Lin (✉)
Veterinary Medicine Research and Development, Pfizer Animal Health, Zaventem, Belgium
e-mail: Dan.Lin2@pfizer.com

W. Talloen · L. Bijnens · H.W.H. Göhlmann
Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium
e-mail: WTALLOEN@its.jnj.com; LBIJNENS@its.jnj.com; HGOEHLMA@its.jnj.com

D. Amaratunga
Biostatistics and Programming, Janssen Pharmaceutical Companies of Johnson & Johnson, Raritan, NJ USA
e-mail: DAMARATU@its.jnj.com

R. Straetemans
Ablynx NV, Zwijnaarde, Belgium
e-mail: Roel.Straetemans@ablynx.com

D. Lin et al. (eds.), *Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R*, Use R!, DOI 10.1007/978-3-642-24007-2_1,

effect profile that is acceptable within the dose range that demonstrates a high level of target activity.

Dose-response experiments have a simple concept. The compound of interest is administered at several doses to a biological sample (a cell line, an animal model, a human volunteer, or a patient) and the response is measured. Dose-response experiments allow researchers to assess the relationship between the dose (amount, concentration) of a drug and the response observed. Despite the conceptual simplicity, however, the practical analysis is much more complicated. First, a response can change dose-dependently in a lot of different ways, and many of dose-response relationship are complex and nonlinear. Second, it is difficult to choose an appropriate response measure. Often one may want to investigate even more than one response. This is because a treatment will generally lead to multiple biological reactions, and one needs to try to disentangle direct from indirect responses and desired from undesired effects.

1.2 Dose-Response Modeling

Dose-response models aim to describe the dependency of a specific response on dose. Dose quantifies the amount of drug the subject is exposed to. Most commonly it measures the weight of the chemical compound and is expressed in absolute terms such as milligram (mg) or grams (g) when dealing with clinical studies or in relative terms such as mg/kg when dealing with in vivo animal experiments. The dose as such, however, is not the direct cause for the response or response profile. In fact, in most cases, the molecular concentration at the site of action drives the response to a drug. This information is seldom readily available or even measurable. A precursor for this information is the drug concentration in the blood. Therefore, whenever drug concentrations in plasma are available, dose can be easily, and arguably should be, replaced by concentration. Since this information is not always at hand, dose is a good alternative. There are methods which under certain circumstances, e.g., knowledge of the pharmacokinetic profile of the compound under study, can be used to work with concentration instead of dose when no plasma samples are available (Jacqmin et al. 2007; Jacobs et al. 2010). A detailed description of these techniques is beyond the scope of this book.

Response can basically be any observation of interest that can be measured. Examples range from continuous data, e.g., body weight data, to binomial data where the presence of an event is observed (yes or no) or multinomial data where multiple levels are possible such as different pain scores or types of adverse events. Although the goal of dose-response modeling for each possible type of response is the same, finding an optimal dose, the methodology behind it will differ greatly from one type to the other. The main focus of the book is on continuous data and in particular the evolution of the mean gene expression with respect to dose.

Ruberg (1995a,b) and Chuang-Stein and Agresti (1997) formulated four main questions usually asked in dose-response studies: (1) Is there any evidence of

the drug effect? (2) For which doses is the response different from the response in the control group? (3) What is the nature of the dose-response relationship? and (4) What is the optimal dose? We can answer these questions either by testing for monotone trend of the response with respect to dose or by modeling the dose-response relationship. In both cases, our underlying assumption is that the true relationship between the dose and the response is monotone. In some applications, the underlying assumption of monotonicity is not appropriate, and other non-monotone order-restricted profiles such as the simple tree order, unimodal partial order (umbrella profiles), and cyclical patterns should be considered. For a discussion about monotonicity issues within the dose-response setting, we refer to Cooke (2009) and Louis (2009).

Dose-response modeling refers to implementing a mathematical representation of some true and unknown relationship. Dose-response models can be classified as being empirical or mechanistic in nature. An empirical model, such as the four-parameter logistic model, discussed in Chaps. 4 and 14, serves to adequately describe the observed pattern between a dose and a response without giving an understanding of the underlying biological process. In other words, the parameters present in the model do not represent biological processes. A mechanistic model on the other hand uses mechanistical pathways to explain the observed pattern. In this book, we focus on the first type of models.

1.3 Dose-Response Microarray Experiments

Now the genome of man and other species have been completely sequenced, we enter the so-called "post-genomic era" that concentrates on harvesting the fruits hidden in the genomic text (Lengauer 2001). The advent of biotechnologies such as microarrays allows us to do so by effectively measuring the activity of an entire genome at once under different conditions. The wealth of biological information of this procedure presents immense new opportunities for developing effective therapies. History has taught us that 30–40% of experimental drugs fail because an inappropriate biological target was pursued (Butcher 2003). The major impact of genomic information may therefore be to reduce this biological failure rate by earlier definition of drug targets related to disease susceptibility or progression. This becomes clearer when one reflects about what a "drug-target" actually is. A drug target is a relatively vague term referring to any number of biological molecular classes (proteins, genes, RNA, sugars, ...) that are "druggable". To be druggable, a target needs to be accessible to putative drug molecules and bind them in such a way that a beneficial biological effect is produced. With microarrays and other high-content screening tools, a wide array of target identification and validation technologies becomes available. Genomics, transcriptomics (i.e., gene expression profiling), and proteomics allow researchers to study many of these drug targets in an unprecedented high-content way. They allow researchers to monitor and discover the biological effects of a potential drug. In summary, the availability of

the human genome sequence represents an exciting advance for the development of novel treatments. When combined with high-content screening methods such as microarrays, the success rate for experimental drugs can be expected to improve (Butcher 2003). The major issue with microarray data analysis is the curse of high dimensionality. Because so much information is gathered on biological activity, it becomes a challenge to find the relevant information in the haystack of irrelevant information. One runs the risk of missing the interesting results in a mass of false positive findings.

Microarray dose-response experiments allow researchers to study the relationship between the dose of a drug and the activity of an entire genome at once. It combines the information wealth of microarrays with the benefits of dose-response studies. Their combined use yields two additional advantages. First, the proportion of false positive findings will be substantially reduced as more information on the entire dose-response profile is collected. False positive genes identified by a one-dose treatment study are easier to unmask in multiple-dose studies when the gene has an unrealistic dose-response relationship. Second, genes within the same biological sample may respond differently to drug dose. One therefore wants to investigate more than one gene. How many exactly is difficult to say, but in a discovery phase, it is typically the more the better. In early stages of drug development, one indeed tries to explore as many potential effects of the drug as possible. A microarray dose-response experiment studies the entire genome at once, and is therefore an ideal tool to elucidate variation in dose-dependency of a treatment across all genes and all known pathways.

Although analysis of gene expression data is the main focus of the book, the discussion about microarray technology is beyond the scope of this book. We refer to Amaratunga and Cabrera (2003) and Göehlmann and Talloen (2009) for an elaborate discussion about the microarray technology and topics related to the analysis of microarray data.

1.4 The Book Structure

The general structure of this book is shown in Fig. 1.1. Although the main part of this book is devoted to the specific setting of microarray dose-response experiment, we introduce the main concept of dose-response modeling in the first part of the book. Estimation under order restrictions and inference are discussed in Chaps. 2 and 3, while parametric nonlinear modeling of dose-response data is described in Chap. 4. The methodology discussed in these chapters is introduced in a general setting, and materials for these chapters are used throughout the second part of the book.

The second part of the book starts with an introduction to dose-response microarray experiments and their specific data structure in Chap. 5, in which the case studies are introduced as well. The analysis of microarray data introduces the challenge of multiple testing. A general guidance for the multiple testing problem in a microarray

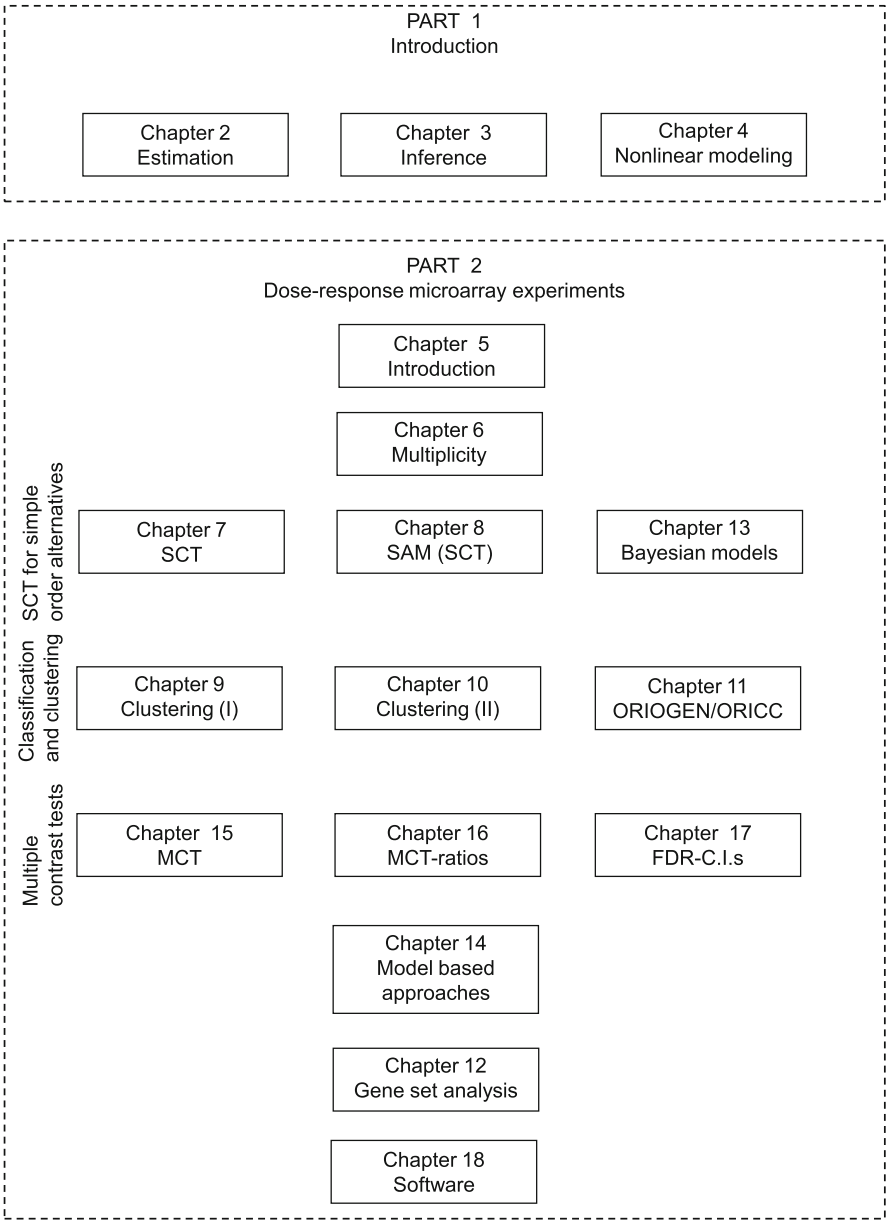

Fig. 1.1 The book structure

setting is given in Chap. 6. We discuss several methods and adjusting procedures such as Bonferoni and Holm's procedures for controlling the family-wise error rate and Benjamini and Hochbergh's procedure and the significance analysis of microarrays (SAM) for controlling the false discovery rate (FDR). Chapters 7, 8,

and 13 are devoted to order-restricted inference in the dose-response microarray setting. We discuss several inference procedures for detecting genes with monotone dose-response relationships such as permutation tests, the SAM for dose-response data, and Bayesian approaches. More advanced inferential topics within the dose-response microarray setting are given in Chaps. 15–17 in which we discuss the topics of multiple contrast tests, ratio tests, and FDR-adjusted confidence intervals. Three chapters in the book are devoted to methods which can be used in order to interpret the results obtained from the inference step. Chapters 9 and 10 present methods for classification and clustering of dose-response curves which can be applied after an initial inference step (discussed in Chaps. 7 and 8). In Chap. 11, we relax the assumption of monotonicity and discuss order-restricted dose-response relationships which are not necessarily monotone such as the unimodal partial order and simple tree order. This setting is further discussed in Chap. 15 in the context of multiple contrast tests. Chapter 12 focuses on the interpretation of the genes detected using the gene set analysis based on Gene Ontology library. As mentioned above, we discuss the general concept of parametric dose-response modeling in Chap. 4 in the first part of the book. In Chap. 14, we focus on parametric modeling of dose-response microarray data. Note that in contrast to other chapters in the second part of the book, the aim of the analysis presented in Chap. 14 is not to detect genes with significant monotone trend, but to perform a secondary analysis in which characteristics of the dose-response relationship are investigated in more details.

The analysis presented in the book is done using several R packages. The methodology discussed in Chaps. 7 and 8 is implemented in the R package `IsoGene` and its graphical user interface `IsoGeneGUI`, which are developed in line with the book. A detailed illustration of the `IsoGeneGUI` is given in Chap. 18, while the use of the `IsoGene` package is illustrated in most of the book chapters. Throughout the book, various R packages are applied for specific settings, which include `nlme`, `multtest`, `ORCM`, `MLP`, `BRUGs`, `MCPMod`, `multcomp`, and `mratios`. Our working assumption is that the readers of this book have a basic knowledge of R, and therefore, complete working examples are provided for the data analysis presented in the book. Readers with limited knowledge of R who wish to perform the analyses presented in the book can do so easily by using the `IsoGeneGUI` which does not require knowledge of R syntax.

1.5 Notation

Throughout the book, we denote d_i as the ith dose level. We use $\mu(d_i)$ and μ_i for the mean gene expression at the ith dose level. Isotonic means at the ith dose are denoted as $\hat{\mu}(d_i)^*$ and $\hat{\mu}_i^*$. Unless specified otherwise, all the models presented in the book are gene specific. In order to simplify notation, we drop the index for the gene. R code is presented in the following way:

```
> # Example of R code.
> age = c(8, 8, 8, 10, 10, 10, 12, 12, 12,14, 14)
> size = c(21,23.5,23,24,21,25,21.5,22,19,23.5,25)
```

Note that, the R code for the analysis is presented as complete as possible. Due to space limit and the length of the code, parts of the code is omitted. The complete R code can be downloaded from the website of the book at:

http://www.ibiostat.be/software/IsoGeneGUI/index.html.

References

Amaratunga & Cabrera, J. (2003). Exploration and Analysis of DNA Microarray and Protein Array Data, New York: John Wiley

Barlow, R.E., Bartholomew, D.J., Bremner, M.J., & Brunk, H.D. (1972). *Statistical inference under order restriction*. New York: Wiley.

Butcher, S. (2003). Target discovery and validation in the post-genomic era. *Neurochemical Research, 28*, 367–377.

Chuang-Stein, C., & Agresti, A. (1997). Tutorial in biostatistics: A review of tests for detecting a monotone dose-response relationship with ordinal response data. *Statistics in Medicine, 16*, 2599–2618.

Cooke, R.M. (Ed.). (2009). *Uncertainty modeling in dose-response*. New York: Wiley.

FDA, U. (2004). *Innovation or stagnation: Challenge and opportunity on the critical path to new medical products*. Silver Spring: US Food and Drug Administation. New York: Wiley

Goehlmann, H., & Talloen, W. (2009). *Gene expression studies using Affymetrix microarrays*. Boca Raton: Chapman & Hall/CRC. New York: Wiley

Jacqmin, P., Snoeck, E., van Schaick, E., Gieschke, R., Pillai, P., Steimer, J.-L., et al. (2007). Modelling response time profiles in the absence of drug concentrations: Definition and performance evaluation of the KPD model. *Journal of Pharmacokinetics and Pharmacodynamics, 34*, 57–85.

Jacobs, T., Straetmans, R., Molenberghs, G., Bouwknecht, A., & Bijnens, L. (2010). Latent pharmacokinetic time profile to model dose-response survival data. *Journal of Biopharmaceutical Statistics, 20*(4), 759–767.

Lengauer, T. (2001). Computational biology at the beginning of the post-genomic era. In *Informatics—10 years back. 10 years ahead* (Vol. 355). Heidelberg: Springer. New York: Wiley.

Louis, T.A. (2009). Math/Stat perspective in Chapter 2: agreement and disagreement, in Cooke, R.M. (Ed.). (2009). *Uncertainty modeling in dose-response*. New York: Wiley.

Marton, M., DeRisi, J., Bennett, H., Iyer, V., Meyer, M. Roberts, C., Stoughton, R., Burchard, J., Slade, D., & Dai, H. (1998). Drug target validation and idenfication of secondary drug target effects using DNA microarrays. *Nature Medicine, 4*, 1293–1301.

Ruberg, S.J. (1995a). Dose response studies. I. Some design considerations. *Journal of Biopharmaceutical Statistics, 5*(1), 1–14.

Ruberg, S. J. (1995b) Dose response studies. II. Analysis and interpretation. *Journal of Biopharmaceutical Statistics, 5*(1), 15–42.

Sams-Dodd, F. (2005). Target-based drug discovery: Is something wrong? *Drug Discovery Today, 10*, 139–147.

Part I
Dose–Response Modeling: An Introduction

Chapter 2
Estimation Under Order Restrictions

Ziv Shkedy, Dhammika Amaratunga, and Marc Aerts

2.1 Introduction

The basic setting on which we focus in the first part of this book is one in which a response variable Y is expected to increase or decrease monotonically with respect to increasing levels of a predictor variable x which in biomedical applications is usually the dose or concentration of a drug. We assume that the mean response is given by

$$E(Y|x) = \mu(x),$$

where $\mu()$ is an unknown monotone function. The case in which $\mu(x)$ is order restricted but not monotone will be discussed in Chaps. 5, 11 and 15 in the second part of the book. The main problem is that although $\mu()$ is a monotone function, unless the sample size at each design point increases to infinity, neither the observed data nor the estimated means are necessarily monotone. For illustration, consider a linear model with five discrete design points:

$$Y_{ij} \sim N(\mu_i, 1), \quad i = 1, 2, 3, 4, 5, \ j = 1, \ldots, n_i. \tag{2.1}$$

Here, μ_i is the true mean at the each design point, $\mu_i = 5$, 5.5, 6, 6.5, and 7, respectively, and n_i is the sample size at each design point. We generate 10×6 datasets according to model (2.1) with the sample size at each design point equal to 5, 10, 25, 50, 100, and 1,000, respectively. Figure 2.1a shows an example of

Z. Shkedy (✉) · M. Aerts
Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat),
Center for Statistics (CenStat), Hasselt University, Diepenbeek, Belgium
e-mail: Ziv.shkedy@uhasselt.be; marc.aerts@uhasselt.be

D. Amaratunga
Biostatistics and Programming, Jansenn Pharmaceutical Companies of Johnson & Johnson,
Raritan, NJ USA
e-mail: DAMARATU@its.jnj.com

D. Lin et al. (eds.), *Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R*, Use R!, DOI 10.1007/978-3-642-24007-2_2,

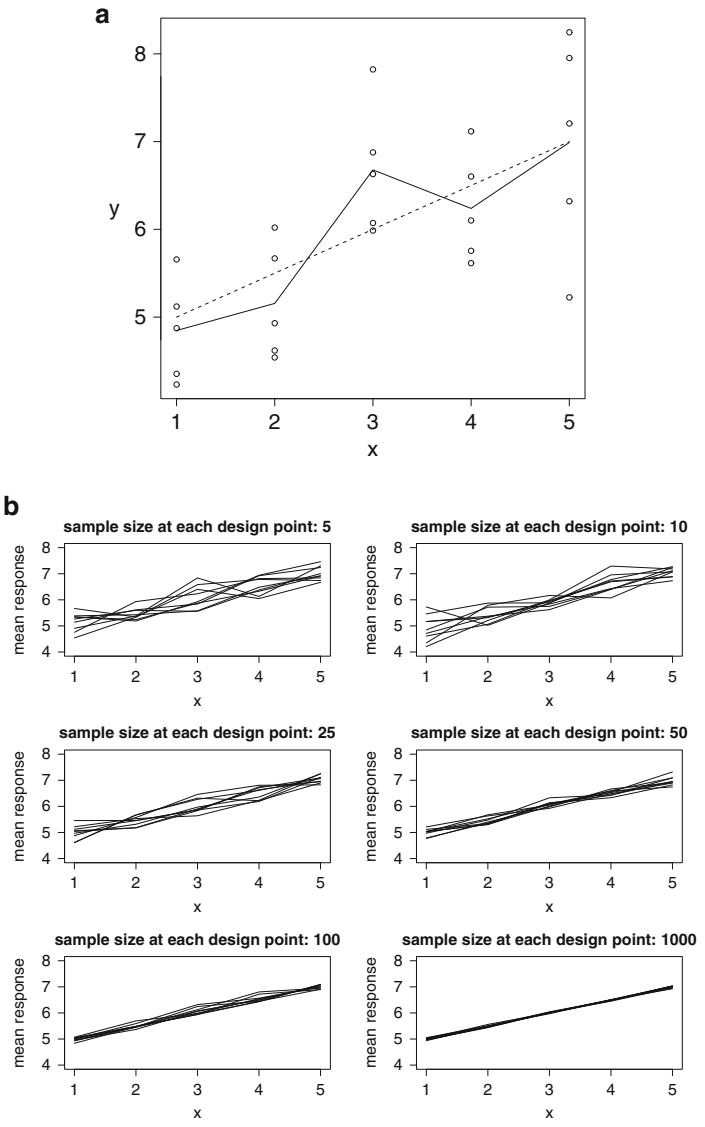

Fig. 2.1 Illustrative example with true mean at each deign point equal to 5, 5.5, 6, 6.5, and 7, respectively. (**a**) Illustrative example of a dataset with five observations at each design point. *Dashed line*: unrestricted means; *solid line*: true (monotone) means. (**b**) Sample means at each design point for ten datasets generated under model (2.1), sample sizes are equal to 5, 10, 25, 50, 100, and 1,000

one dataset with five observations at each design point while Fig. 2.1b shows the estimated means for ten datasets for an increasing number of observations at each design point. Clearly, when sample sizes are relatively small, the observed means are not monotone even though the true means are monotone.

Two questions arise now. The first is how to estimate the mean response under the assumption that the true mean is monotone with respect to x. The second is how to test whether the mean responses in the population are indeed monotone. In this chapter, we focus on the first question and discuss the estimation problem using isotonic regression while the topic of inference under order restrictions will be discussed in Chap. 3.

2.2 Isotonic Regression

According to Barlow et al. (1972), isotonic regression is the statistical theory that deals with problems in which conditional expectations are subject to order restrictions. Let $X = \{x_1, x_2, \ldots, x_n\}$, where $x_1 \leq x_2 \leq \cdots \leq x_n$ denote the finite set of n observed values for the predictor variable and $y_1, y_2, \ldots, y_n$ denote the corresponding observed values for the response variable. Let $\mu(x)$ denote the mean of the conditional distribution, $\mu(x) = E(Y|x)$. Within the framework of linear regression, estimating $\mu(x)$ is typically done by minimizing the least squares criterion in the class of arbitrary linear functions f on X. However, it might be assumed or known that $\mu(x)$ is nondecreasing in x, that is, isotonic with respect to the simple order on X. In that situation, isotonic regression refers to minimizing the least squares criterion in the class of isotonic functions f on X, i.e. f is isotonic, if $x_i \leq x_j, i \neq j$ implies that $f(x_i) \leq f(x_j), i, j = 1, \ldots, n$. Barlow et al. (1972) proposed using the "pool adjacent violators algorithm" (PAVA), where successive approximation is used to isotonize the minimizer of the least squares criterion. Robertson et al. (1988) defined isotonic regression as follows.

Let g be a given function of X. A function $g^\star$ on X is an isotonic regression of g with weight w if and only if $g^\star$ is isotonic and $g^\star$ minimizes

$$\sum_{x \in X} [g(x) - f(x)]^2 w(x), \tag{2.2}$$

in the class of all isotonic functions f on X. The PAVA was proposed by Barlow et al. (1972) and Robertson et al. (1988) in order to minimize (2.2) subject to the constraint $f(x_j) \leq f(x_i)$ for $x_j \leq x_i$.

2.2.1 The PAVA

Let us focus again on an experiment in which the predictor variable X has $K+1$ discrete levels and the response variable has n_i replicates at each level of the predictor variable. Hence, at each design point, the observed data consists of the pairs $\{(x_i, y_{ij})\}$, $i = 0, 1, \ldots, K$, $j = 1, \ldots, n_i$. Without loss of generality, we assume $x_0 \leq \cdots \leq x_K$. Denote the maximum likelihood estimate of $\mu(x_i)$ by $\hat{\mu}(x_i)$.

Suppose i^* is the first index for which $\hat{\mu}(x_i) \geq \hat{\mu}(x_{i+1})$, i.e., the first index for which a "violation" of monotone behavior is observed. The PAVA states that these values need to be "pooled." In other words, $\hat{\mu}(x_i)$ and $\hat{\mu}(x_{i+1})$ are both replaced by the weighted mean:

$$\hat{\mu}(x_i, x_{i+1}) = \frac{n_i \hat{\mu}(x_i) + n_{i+1}\hat{\mu}(x_{i+1})}{n_i + n_{i+1}}.$$

The algorithm proceeds by recursively checking monotone behavior and by pooling if necessary and finally stops if monotonicity is achieved.

2.2.2 Example 1: Size of Pituitary Fissure (Robertson et al. 1988)

Robertson et al. (1988) discussed a dental study in which the size of the pituitary fissure was measured for groups of girls at age 8, 10, 12, and 14. The raw data are shown in Fig. 2.2a. The underlying assumption is that the size of the pituitary fissure increases with age. However, as can be seen clearly in Fig. 2.2a, the observed mean at each age group is not monotone with age since the observed mean at age 12 (20.8333) is smaller than the observed mean at age 10 (23.3333).

```
> age = c(8, 8, 8, 10, 10, 10, 12, 12, 12,14, 14)
> size = c(21,23.5,23,24,21,25,21.5,22,19,23.5,25)
> #sample means at each age group
> msize<-tapply(size,as.factor(age),mean)
> par(mfrow=c(2,2))
> plot(age,size)
> lines(unique(age),msize,lty=2)
> title("a:row data")
> msize
       8       10       12       14
22.50000 23.33333 20.83333 24.25000
```

In the first step of the PAVA, we pool together the means of the second and third age groups,

$$\hat{\mu}(10, 12) = \frac{n_{10}\hat{\mu}(10) + n_{12}\hat{\mu}(12)}{n_{10} + n_{12}}.$$

```
> msize1<-msize
> msize1[2:3]<-(3*23.33333+3*20.83333)/6
> msize1
       8       10       12       14
22.50000 22.08333 22.08333 24.25000
> plot(age,size)
> lines(unique(age),msize,lty=2)
> lines(unique(age),msize1)
> title("b:step 1")
```

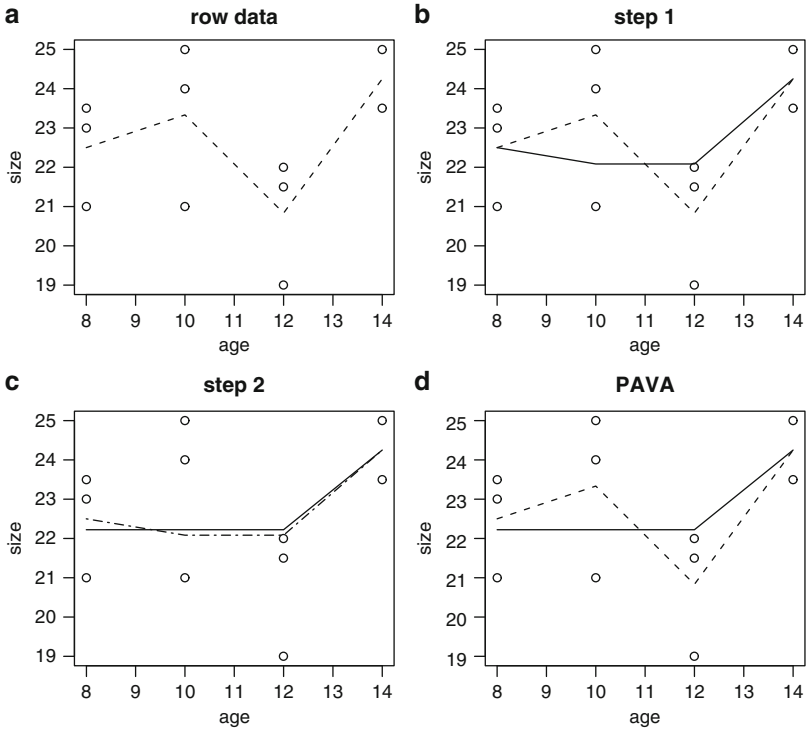

Fig. 2.2 Isotonic regression for the pituitary fissure example from Robertson et al. (1988). Panels (**a**)–(**c**): isotonic regression step by step. Panel (**d**): isotonic regression using the `pava()` function. *Dashed line*: unrestricted means

The result is presented in Fig. 2.2b, and we can see that after the first pooling, the mean in the first age group is higher than the pooled mean of the second and third groups, and therefore, a second pooling is needed,

$$\hat{\mu}(8,10,12) = \frac{n_8\hat{\mu}(8) + (n_{10}+n_{12})\hat{\mu}(10,12)}{n_8 + n_{10} + n_{12}}.$$

```
> msize2<-msize1
> msize2[1:3]<-(3*22.5+6*22.08333)/9
> msize2
       8       10       12       14
22.22222 22.22222 22.22222 24.25000
> plot(age,size)
> lines(unique(age),msize1,lty=2)
> lines(unique(age),msize2)
> title("c:step 2")
```

Figure 2.2c shows that after the second pooling, the isotonic means are monotone. Isotonic regression can be fitted using the R function `pava()`. A general call of the function in the `Iso` package has the form