Modelling High Dimensional Dose-Response Data

Mgr. Martin Otava, MSc

Promotor: Prof. dr. Ziv Shkedy

Co-Promotor: Dr Adetayo Kasim

Co-Promotor: Prof. dr. Willem Talloen

Lubošovi, Evě, Radkovi a babičce Jarče. E à Renata.

"Bože, jak jednoduchý recept na štastný život - to, co děláme, dělat z lásky k věci."
Karel Čapek

Acknowledgements

I have met so many great people during past four years, both among the colleagues at various places, as well as in personal life. Therefore, the acknowledgement would never contain all of the things I am grateful for, nor the people I would like to thank. It will be just the brief and incomplete summary.

I was very lucky to have Ziv Shkedy as the supervisor, because his working style was just perfect for me. Especially, I am very grateful for his availability to set meeting whenever I felt it is needed and also for his valuable advices on all the different aspects of the PhD. He was also very supportive in "side-projects" that extended my experience broadly, namely teaching in Ethiopia and teaching acivities for Master students in Hasselt. Besides being great supervisor, Ziv is great person in general and it is indeed very easy to study and work hard, if you can consider your supervisor being your friend. I am also grateful to Adetayo Kasim for becoming my co-supervisor, although being very busy with building his department at Durham University, and being of great source of inspiration, and to Dan Lin who kept in touch after leaving the department, for good advices and friendship.

I learned a lot during my visits to Beerse and I was very lucky to be able to get experience from industry as well as academia. It was not only about direct interaction with the great professionals there, but also about interaction with non-statisticians, huge amount of presentations that we had to give and access to the network of specialists and opportunity to learn from them (just for one example is unforgettable presentation about R by José Pinheiro). My main thanks goes to Willem Talloen and Luc Bijnens whom I collaborated most extensively and who taught me a lot. However, I would like to extend my acknowledgement to the whole Nonclinical statistics team for providing such a stimulating environment, as well as people collaborating on QSTAR and ExaScience projects.

I was honored to have very good collaborators for various topics of the thesis. Dani Yekutieli and Frank Bretz for permutation test of BVS, Ludwig Hothorn and Daniel Gerhard for model selection problems, Geert Verheyen for pathway analysis and toxicogenomics. Also, at our department, there were so many nice people: JOSS board, office mates from B2 and E101, thanks for good times! Special thanks goes indeed to Martine and Hilde for being incredibly helpful and efficient at any time!

Nolen, salamat! You were the best colleague ever and also great friend to me! I wish you all the best wherever you go and I hope to visit you in Philippines one day.

Eva and Jimmy, thanks a lot for making me busy at the weekends! I admire your attitude and I still do not fully comprehend, how can you make all the that stuff while full time working. Koen, thanks for sharing all that different events all over the year and for jogging! Yovanna, muchas gracias por ser gran amiga! Especially at the beginning, when I missed my family a lot, visiting you, Miguel and Sulay always felt like coming home. Emanuele, Fortunato, Donato, Consu, I will never forget the longest Easter dinner in my life nor the great evening parties at Nierstraat and salsa in Genk! You made my stay here so much more pleasant! As well as many others: thank you Caro, Kim, Sammy, Chella, Kathy, Ambily, Tanya, Wibren, Ariel, Pia, Izabela, Nikolina, Wiebke, Farnoosh, Trishanta, Yimer and many more for being such great friends.

Rád bych poděkoval všem přátelům doma, kteři na mě nezapomněli a zůstali v kontaktu. Pokoušel jsem se původně o jmenný seznam, ale začínal být neúnosně dlouhý a stejně bych musel opomenout spoustu lidí. Veřte mi proto, ze jsem měl radost z každého hovoru na Skypu, emailu a že jsem si nikdy nemohl stěžovat, ze bych v Čechách neměl co dělat. Občas byla výzva spojení udržet a se spoustou z vás jsem mluvil a viděl se mnohem méně, než bych si býval přál. Na druhou stranu, nevěřím, že bych to tady dostudoval, kdybych někdy získal pocit, že ztrácím kontakt s vámi všemi. Doufám, že nám to vydrží i nadále a že bude dost příležitostí se vídat. Samozřejmě, jste všichni zvaní na návštěvu! Speciální poděkování pro Čendu, Zdendu a Jardu, což snad nemusím nijak vysvětlovat. Hynkovi za tu hromadu hovorů a Kamče a Petrovi (nejen) za skvělou společnou dovolenou. Dalši velké poděkování patří všem, co se podílejí na letním táboře, ať už na straně organizátorů či účastníků, za tu úžasnou atmosféru a to, jak moc jsem si tam vždycky odpočinul. Dolly, Honzo, Martine, Jardo, Nathe, Vláďo a Kiki, díky za sdílení chatky ve všech těch různých letech, byla to paráda.

Na závěr patří poděkování mé rodině. Děkuji za podporu, rady a starost za všech okolností! Přijet sem mi dalo hodně, ale stejně tak jsem toho doma spoustu propásnul. Mám vás moc rád a vždycky tu pro vás budu, ať budu jakkoli daleko!

Finalmente, muito obrigado, meu amor. Para tudo. Eu não iria ter sucesso sem você. Te amo muito!

Publications

The materials presented here are based on the following publications and reports:

Manuscripts and book chapters

Kasim, A., Van Sanden, S., **Otava, M.**, Hochreiter, S., Clevert, D.-A., Talloen, W., Lin, D. (2012) δ-clustering of Monotone Profiles. *In* Lin, D., Shkedy, Z,. Yekutieli, D., Amaratunga, D., Bijnens, L. (ed.), *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R*, Springer, Berlin, pp. 193-214.

Otava, M., Shkedy, Z., Kasim, A. (2014) Prediction of Gene Expression in Human Using Rat in Vivo Gene Expression in Japanese Toxicogenomics Project. *Systems Biomedicine*, 2:e29412. DOI:10.4161/sysb.29412.

Otava, M., Shkedy, Z., Lin, D., Göhlmann, H. W. H., Bijnens, L., Talloen, W., Kasim, A. (2014) Dose-Response Modeling Under Simple Order Restrictions Using Bayesian Variable Selection Methods. *Statistics in Biopharmaceutical Research*, 6(3), 252-262. DOI: 10.1080/19466315.2013.855472.

Otava, M., Lin, D., Shkedy, Z., Kasim, A., Verbeke, T., Pramana, S., Bijnens, L., Göhlmann, H. W. H., Talloen, W. (2015) δ-Clustering of Monotone Profiles for Dose-response Gene Expression Data: The ORCME R Package. *To be submitted*.

Otava, M., Shkedy, Z., Talloen, W., Verheyen, G. R., Kasim, A. (2015) Identification of *in vitro* and *in vivo* disconnects using transcriptomics data. *BMC Genomics, 16, 615. DOI 10.1186/s12864-015-1726-7.*

Otava, M., Shkedy, Z., Lin, D., Pramana, S.,Verbeke, T., Haldermans, P., Hothorn, L. A., Gerhard, D., Kuiper, R., Klinglmueller, F., Kasim, A., (2015) IsoGeneGUI: multiple approaches for dose-response analysis of microarray data using R. *Submitted to R-Journal.*

Otava, M., Lin, D., Shkedy, Z., Bretz, F., Talloen, W., Yekutieli, D., Kasim, A. (2015) Order restricted Bayesian inference under model uncertainty for dose-response experiments. *To be submitted.*

Otava, M., et al (2015) Identification of the Minimum Effective Dose for Normally Distributed Endpoints Using a Bayesian Variable Selection Approach. *To be submitted to Journal of Biopharmaceutical Research.*

Otava, M. (To be published 2016) Patterns Discovery in High Dimensional Problems. *In* Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S., Talloen, W. (ed.), *Applied Biclustering Methods for Big and High Dimensional Data Using R.* Chapman and Hall / CRC.

De Troyer, E., **Otava, M.**, et al (To be published 2016) The BiclustGUI Package. *In* Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S., Talloen, W. (ed.), *Applied Biclustering Methods for Big and High Dimensional Data Using R.*

De Troyer, E., **Otava, M.**, et al (To be published 2016) We R a Community - Including a New Package in BiclustGUI. *In* Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S., Talloen, W. (ed.), *Applied Biclustering Methods for Big and High Dimensional Data Using R.*

Conference proceedings

Otava, M., Kasim, A., Shkedy, Z., Kato, B. S. (2012) Bayesian variable selection method for modeling dose-response microarray data under simple order restrictions. *In* Komárek, A., Nagy, S. (ed.), *Proceedings of the 27nd International Workshop on Statistical Modelling (IWSM)*, pp. 193-214.

Software development

Kasim, A., **Otava, M.**, Verbeke, T. (2014) ORCME: Order Restricted Clustering for Microarray Experiments. R package version 2.0.1. http://CRAN.R-project.org/package=ORCME.

Pramana, S., Lin, D., Haldermans, P., Verbeke, T., **Otava, M.** (2014) IsoGeneGUI: A graphical user interface to conduct a dose-response analysis of microarray data. R package version 2.0.0. http://ibiostat.be/online-resources/online-resources/isogenegui.

Aregay, M., **Otava, M.**, Khamiakova, T., De Troyer, E. (2014) BcDiag: Diagnostics plots for Bicluster Data. R package version 1.0.7. http://CRAN.R-project.org/package=BcDiag.

De Troyer, E., **Otava, M.** (2015) RcmdrPlugin.BiclustGUI: Rcmdr Plugin-in. R package version 0.6.2/r48. http://R-Forge.R-project.org/projects/biclustgui/.

Contents

List of Abbreviations

AIC	Akaike Information Criteria
ANOVA	Analysis of Variance
BH	Benjamini-Hochberg
BY	Benjamini-Yekutieli
BIC	Bayesian Information Criteria
BVS	Bayesian Variable Selection
cFDR	Conditional False Discovery Rate
CRAN	The Comprehensive R Archive Network
DIC	Deviance Information Criteria
DILI	Drug Induced Liver Injury
EGF	Epidermal Growth Factor
FARMS	Factor Analysis for Robust Microarray Summarization
FDA	Food and Drug Administration
FDR	False Discovery Rate
FWER	Family-wise Error Rate
GO	Gene Ontology
GORIC	Generalized Order Restricted Information Criterion
GUI	Graphical User Interface
GVS	Gibbs Variable Selection
HSD	Honest Significant Difference
HESCA	Human Epidermal Squamous Carcinoma
IC	Information Criterion
I/NI	Informative/Non-Informative
KEGG	Kyoto Encyclopedia of Genes and Genomes
LRT	Likelihood-ratio Test
MCMC	Markov Chain Monte Carlo
MCT	Multiple Contrast Test

MED	Minimum Effective Dose
NMR	Nuclear magnetic resonance
ORCME	Order Restricted Clustering for Microarray Experiment
ORIC	Order Restricted Information Criterion
ORICC	Order Restricted Information Criterion-based Clustering
PAVA	Pool Adjacent Violators Algorithm
pWSS	Penalized Weighted Sum of Squares
Rcmdr	R Commander
RNA	Ribonucleic acid
RSS	Residual Sum of Squares
SAM	Significance Analysis of Microarrays
SCT	Single Contrast Test
SD	Standard Deviation
SSVS	Stochastic Search Variable Selection
TGP	Japanese Toxicogenomics Project

Chapter 1

Introduction

The work presented in this thesis is focused on dose-response relationships in a broad sense. The proposed methods can be applied to any experiment with an ordered exposure (such as time, dose, age, temperature, etc.) in which the response is continuous such as drug development, ecological or economical studies. The natural ordering of the exposure variable is the main characteristics of the experiment.

The methods discussed in this thesis lie on the border of biostatistics and statistical bioinformatics. Although the focus is on methodological development in general, the research has been conducted with high dimensional data as main application area in mind. Upscaling the analysis to a high dimensional data implies that the analysis should be carried over from the setting of a single experiment to the case in which thousands of experiments under the same design are performed simultaneously. In such a case, it is impossible to evaluate each experiment using visualization techniques or multiple models fitting as it is typically done for a single experiment. From that reason, automated methods which offer clear decision rules (and preferably account for model uncertainty) should be preferred. Indeed, in case of thousands of experiments, multiplicity corrections should be taken into account in order to provide protection against false findings, caused by chance.

The thesis consists of three parts. The first part is focused on the methodological developments while the other two parts are focused on applications within the bioinformatics domain. The connection between the three parts is the data structure and the modelling approaches, i.e dose-response experiments and an order restricted modelling approach.

In the first part of the thesis, we present a state-of-the-art statistical framework in a generic way so the methods are applicable in a general context. The aim is to elaborate on the theoretical foundations as well as on the empirical evaluation of the proposed methodology. An investigation of the methods' properties is done through extensive simulation studies within various settings. The focus of the first part is placed on the order restricted Bayesian variable selection (BVS) modelling framework. The advantage of the BVS approach is that the method allows for simultaneous estimation and model selection, while adjusting for model uncertainty. Note that variable selection refers to selection of which doses have an effect on response instead of selection of independent variables to be included in the model. Analogously, model selection is related to selection of underlying dose-response profile. In the first part of the thesis, the BVS method is extended to allow inference using resampling based techniques. Hence, it offers an unified framework for order restricted data analysis without necessity to apply any post hoc methodology. Moreover, its Bayesian nature allows for incorporation of prior scientific knowledge whenever available.

The BVS method is discussed over several chapters in the first part of the thesis. Chapter 2 provides a detailed introduction to the topic. Chapter 3 introduces a resampling based inference procedure within the BVS framework. Model selection and the determination of the minimum effective dose (MED) are the main subjects of Chapter 4. The MED is an example of importance of model selection framework. Any other quantities based on the dose-response profile can be computed in analogously, based on selected model or using model averaging, taking into account model uncertainty. The robustness of the inference, model selection and estimation procedures against the specification of prior distributions is investigated in Chapter 5. In addition, model complexity is defined and its properties within BVS modelling framework are analyzed in Chapter 5, as well. Finally, Chapter 6 describes in detail the simulations studies conducted in order to investigate the performance of the methods discussed in the previous chapters.

The second part of the thesis focuses on the analysis of one database. The target of this part is developing a data analysis workflow in order to analyze complex multisource data sets and to extract knowledge out of them. Rather then developing a new methodology, the aim in the second part is to use known and validated methods in a novel and efficient way. Although the focus is on the analysis of one particular database, the workflow can be generalized further for similar problems in a broader sense within the research domain.

The case study analyzed in the second part is a large toxicogenomics database. Two analysis frameworks are presented, each of them is focused on the translational research from a different point of view. In the first analysis, the primary interest is the identification of genes with similar dose-response profiles in two related data sets. In contrast, the second analysis focuses on the identification of genes showing strong discrepancies between two data sets. Both groups of genes are of interest under varying research questions and their identification pose different statistical problems. Therefore, methods used in the analyses range from order-restricted dose-response modelling techniques to fractional polynomial models that relax the monotonicity assumption. We used biclustering and visualization methods to explore the data and to reveal interesting data patterns. Strong emphasis is given to the interpretation of the results and to the identification of local patterns in the output of the analysis. It is important to realize that both analyses represent exploratory tools starting from general research questions and leading to sets of genes. These resulting genes may have desired properties or relationships with the response, but due to the exploratory nature of the algorithms, scientific knowledge needs to be applied and further validation experiments need to be conducted to evaluate the obtained findings. The case study demonstrates how statistical techniques can be applied to large multisource data and how to interpret the results.

The analysis of the toxicogenomics project is presented in two chapters. In Chapter 7, we search for the genes translatable between rat *in vivo* and human *in vitro* data. In contrast, in Chapter 8, genes disconnected in their effects across platforms, i.e. rat *in vitro* and rat *in vivo*, are identified.

Within the research work related to the PhD project an important effort was to provide data analysis tools for the scientific community. We focused on software development in R (R Core Team, 2014) for its high quality, wide availability and open access environment. In the third part of the thesis we present two R packages. The first R package, ORCME, presented in Chapter 9, performs an order restricted clustering for microarray experiments, the framework that is typically used in the exploratory data analysis stage. The package is available in the Comprehensive R Archive Network (CRAN, Hornik, 2012) repository and its target users are scientist with at least basic experience with R. The second package IsoGeneGUI introduced in Chapter 10 is implemented as a Graphical User Interface and is available in Bioconductor to a wider community of scientists working on biostatistical problems. The point-and-click nature of the package makes it usable to scientists with very limited experience with R.

Chapter 11 concludes the thesis with summary of the work and discussion of possible extensions and further topics for further research.

1.1 Case studies

Several data sets, used in the first part of the thesis, are presented in this section. All datasets were used to illustrate different methods discussed in the first part and demonstrate their proprieties. All the data sets are publicly available.

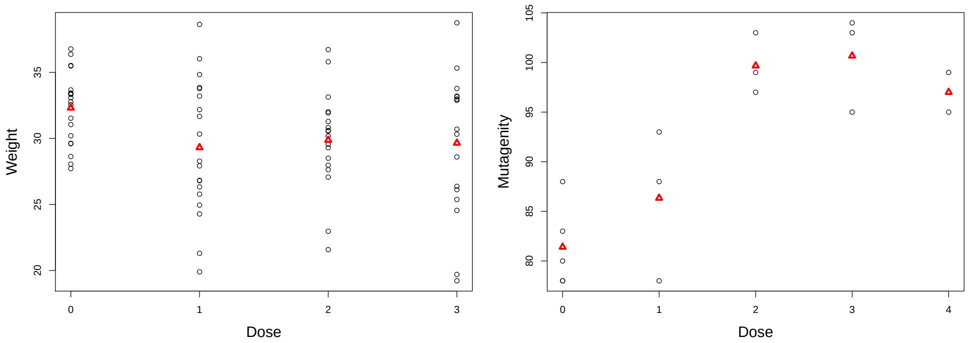


Figure 1.1: *Left panel: The Litter data set. Right panel: The Ames data set. Triangles represent dose-specific means.*

1.1.1 The Litter data

The Litter data set (Westfall and Young, 1993) is available as part of the R (R Core Team, 2014) package `multcomp` (Hothorn *et al.*, 2008). It contains data about pregnant mice that were divided into four groups and the compound in four different doses was administered during pregnancy. For a placebo, 20 mice were used, for active doses 19, 18 and 17 mice, respectively. The litters were evaluated for birth weights. We focus on relationship between the birth weight and the dose. For the Litter data set, the null hypothesis of no dose effect is tested against the nonincreasing alternative in order to detect toxicity effects due to the used drug. The data set is shown in the left panel of Figure 1.1.

1.1.2 The Ames data

The Ames data set (Bretz and Hothorn, 2003) contains the data about a mutagenicity level of a compound, measured under increasing doses of the compound with the first dose being a control (placebo). The mutagenicity is reflected by an increasing relationship between dose level and frequency of visible colonies among plated salmonella bacteria. Dose level is used as a covariate and a frequency of colonies as a response. Although we suspect very high doses to lower number of microbes due to toxicity, in the following analysis we assume only the nondecreasing profile. More detailed information about the data can be found in Ames *et al.* (1975). Five observations are available for a placebo and three for each of four active doses. The data set is shown in the right panel of Figure 1.1.

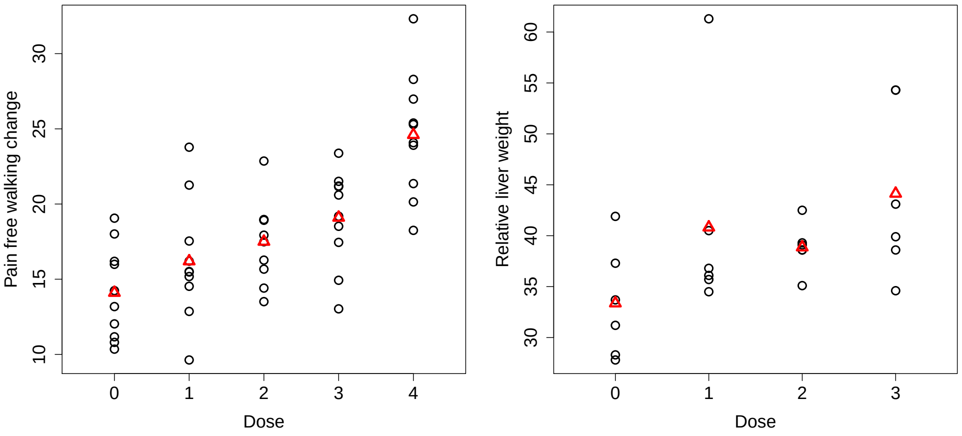


Figure 1.2: *Left panel: The Angina data set. Right panel: The Toxicity data set. Triangles represent dose-specific means.*

1.1.3 The Angina data

The Angina data set (Westfall *et al.*, 1999, p. 164) represents dose-response study of a drug to treat angina pectoris. The response is the duration (in minutes) of pain-free walking after treatment relative to the values before treatment. Four active doses were used together with a control dose with placebo only. Ten patients per dose were examined. Large values indicate positive effects on patients. The data were used in Kuiper *et al.* (2014) and are available under the name `angina` in the package `mratios` (Djira *et al.*, 2012) of the `R` software. Data set is displayed in left panel of Figure 1.2.

1.1.4 The Toxicity data

The Toxicity data set was introduce by Yanagawa and Kikuchi (2001, p. 320) and recently used by Kuiper *et al.* (2014). It represents results of a chronic toxicity study on Mosapride Citrate (Fitzhugh *et al.*, 1964). Liver weight relative to the body weight was measured for 24 dogs. Three active doses of Mosapride Citrate were used and a control dose was added, six dogs were treated in each group. An increasing response suggests an increasing toxicity of the drug.

1.2 Omics case studies

The data sets presented in previous section could be considered as traditional data sets. They consist of a response variable, some explanatory variables and number of independent observations that allow us to estimate the parameters of interest. All these data sets are outcomes of single experiments. The data presented in this section are outcomes of microarray experiments, belonging to the family of 'Omics' data. It typically comprises thousands of variables of interest while having only dozens of observations and it encompasses several data sources or experiments. The standard framework of estimation is disrupted, since the number of possible parameters far exceeds amount of information in the data. Therefore, the sheer size of the data set is challenging to handle, leading to necessity of dimension reduction techniques, multiplicity corrections and careful interpretation of results. Moreover, integration of results of several experiments bring additional challenges. Additionally, the data sets were often not collected in order to test specific hypothesis of interest.

1.2.1 The HESCA study

The HESCA data set (Bijnens *et al.*, 2012) describes results of a dose-response microarray oncology experiment designed to better understand the biological effects of growth factors in human tumor. Human epidermal squamous carcinoma cell line A431 (HESCA431) was grown and cells were stimulated with the epidermal growth factor EGF at four concentrations (including placebo) for 24 hours. Gene expression levels were measured using GeneChip (Affymetrix). The data set contains 12 arrays, three arrays for each of four dose levels with 16,998 probe sets (we would refer to them as genes for simplicity). For details about methodology and preprocessing including normalization, see Bijnens *et al.* (2012).

1.2.2 The Japanese Toxicogenomics Project

The 'Toxicogenomics Project - Genomics Assisted Toxicity Evaluation system' (TG-GATEs, TGP, Uehara *et al.*, 2010) is a collaborative initiative between Japanese National Institute of Health Science, the National Institute Biomedical Innovation and fifteen pharmaceutical companies. It was completed in 2007 after five years of research and it represents a unique source of information for toxicology and safety studies. It offers a rich source of transcriptomics data related to toxicology, providing human *in vitro* experiments together with *in vitro* and *in vivo* rat experiments (Ganter *et al.*, 2005, Suter *et al.*, 2011, Briggs *et al.*, 2012). Almost 20,000 array of Affymetrix platform were generated for liver tissue both *in vitro* and *in vivo* experiments, at various doses and time point for 131 compounds. The compounds are mainly therapeutic drugs, comprising wide range of chemotypes. The TGP contains four main experiments. Three experiments are performed with independent samples: human *in vitro*, rat *in vitro* and rat *in vivo* experiment. Last experiment contains repeated measures for rats *in vivo* and would not be considered further in this thesis. Also, supportive histopathological, hematological and blood chemistry data, obtained for *in vivo* experiments would not be used further. Several toxicogenomics studies on the TGP data set concentrate mostly on network building for rat *in vivo* (Kiyosawa *et al.*, 2010) or the connection between rat *in vivo* and human *in vitro* transcriptomics signatures, with special interest in drug induced liver injury (e.g. Uehara *et al.*, 2008, Clevert *et al.*, 2012, Otava *et al.*, 2014).

Both rat data sets were created using Affymetrix arrays chip Rat230_2. Six weeks old male Sprague-Dawley rats were used for the experiments. Primary hepatocytes were used for *in vitro* experiment; for *in vivo* experiment, each rat was administered a specific dose of a compound and was sacrificed after a fixed time period. Liver tissue was subsequently profiled for gene expression. For the *in vitro* experiments, a modified two-step collagenase perfusion method was used to isolate liver cells from six weeks old rats. These primary cultured hepatocytes were then exposed (in duplo) to a compound and gene expression changes were investigated at multiple time points. Each compound was tested at four different doses, three active doses and placebo (except three compound that were missing either highest or middle dose). Instead of the numerical value of the dose level, expert classification as 'low', 'middle' or 'high' dose is used. This representation was created to allow comparison of compounds with varying potency (and so different actual value of dose). The experiment was conducted at three (*in vitro*, two, eight and 24 hours) or four different time points (*in vivo*, three, six, nine and 24 hours). Each compound, dose and time point combination was tested on multiple independent biological replicates to evaluate variability: duplicates for *in vitro* and triplicates for *in vivo* experiment. Therefore, in total, we have 24 arrays per compound (two biological replicates, four dose levels, three time points) *in vitro* data set and 48 arrays per compound (three biological replicates, four dose levels, four time points) *in vivo* data set.

The human gene expression was measured on primary hepatocytes using Affymetrix chip HG-U133_Plus_2. The compound were tested on three to four dose levels and two to three time points (two, eight and 24 hours), with two independent biological replicates per combination. Therefore, the compound have 16-24 arrays per compound, in total. Again, the expert classification as 'low', 'middle' or 'high' dose is used. All the compounds have at least 12 arrays, being tested on three dose levels (control, middle and high dose) and two time points (eight and 24 hours).

Additionally, the compounds were classified according their drug-induced liver injury (DILI) potential in human, based on their FDA-approved (Chen *et al.*, 2011). In total 101 compounds had the FDA labeling available, resulting in 41 compounds with high or moderate severity of liver injury, 52 compounds with low severity liver injuries or adverse reactions in liver and only eight compounds with no concern related to DILI.

The whole database, together with additional project TGP 2, is available on website http://toxico.nibio.go.jp/english/index.html.

1.2.2.1 Translatability data

The data set is a subset of the TGP data set consists of 93 compounds that are common in rat *in vivo* and human experiments and have DILI information available. In total, 4,440 Affymetrix microarrays that measured gene expression profiles are available for rats (91 compounds with 48 arrays and two compound with 36 arrays) and 1,116 arrays are available for humans (12 arrays per compound). We consider only genes that are orthologous for rats and humans. Further, we filter the genes using the I/NI calls criterion (Talloen *et al.*, 2007). The preprocessed and filtered data set consists of 4,359 genes. Response is computed as log ratio of the gene expression level against mean of expression levels under control dose (vehicle). The gene expression values are based on FARMS (Hochreiter *et al.*, 2006) summarized data. Although the response of interest is a function of gene expression values, we call it 'gene expression' throughout the thesis, for the sake of simplicity. Example of the data is given in Figure 1.3.

1.2.2.2 Disconnect data

The data set is a subset of the TGP data set and consists of 131 compounds that are in common to rat *in vitro* and rat *in vivo* experiment. Three compounds are not suitable for the analysis due to the absence of the data for one of the dose levels. Therefore, the analysis is applied on 128 compounds, for which there are complete rat *in vivo* and *in vitro* data. Only the last time point (24 hours) was considered for the analysis presented in this data set, because there was much stronger signal across genes expressed at 24 hour than at the earlier time points (Otava *et al.*, 2014).

Eventually, 1,024 arrays (eight arrays per compound) and 1,536 arrays (12 arrays per compound) were used for *in vitro* and *in vivo* experiments, respectively. Using I/NI calls filtering (Talloen *et al.*, 2007, Kasim *et al.*, 2010), 5,914 genes are considered reliable and selected for further analysis. The response variable represents the logarithm of the ratio of the original gene expression level against the mean of the gene expression of observations under the control dose. The gene expression values are obtained through the FARMS

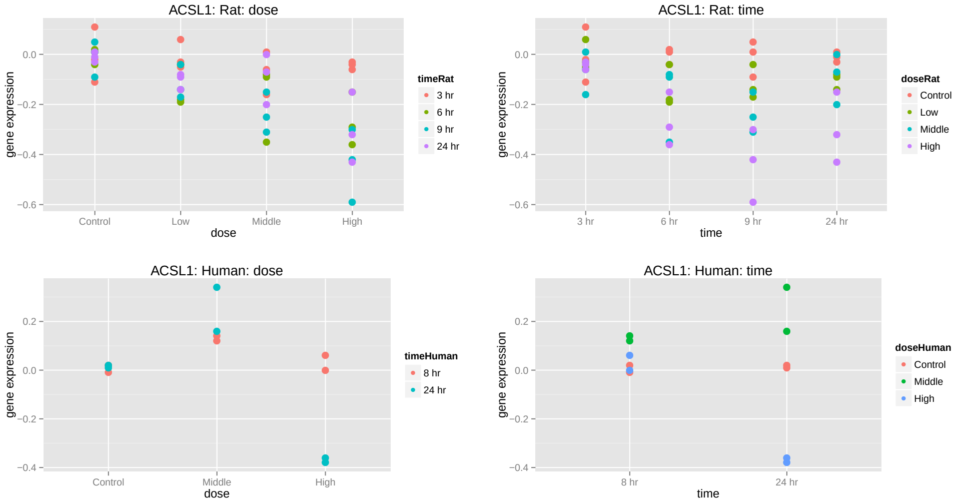


Figure 1.3: *Compound omeprazole and gene Acsl1 in rat and ACSL1 in human, respectively, for* in vitro *experiment. Left and right panels visualize same data. Left panels show for dose-response relationship coloured by time and right panels show time-course data coloured according to dose level.*

summarization method (Hochreiter *et al.*, 2006). Although the response of interest is a function of gene expression values, we call it 'gene expression' throughout the thesis, for the sake of simplicity.

Since only one time point was used, the rat *in vitro* data comprises of eight arrays per compound only (two biological replicates for each of the three active doses and the control dose) and the rat *in vivo* data of 12 arrays per compound (same design, but with three biological replicates per dose level). An example of a dose-response profile of the gene *A2m* within compound sulindac is shown in Figure 1.4.

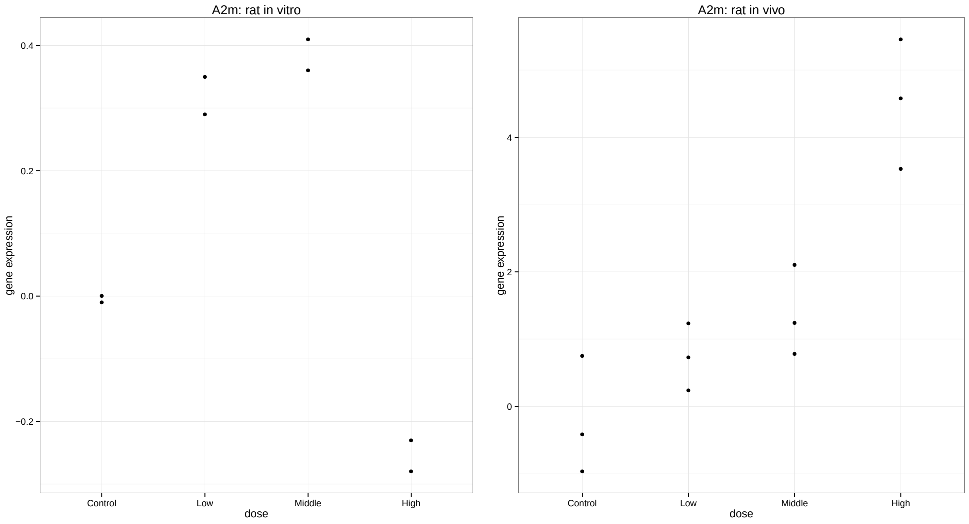


Figure 1.4: *Gene A2m for compound sulindac, last time point only. Left panel:* in vitro. *Right panel:* in vivo.

Part I

Bayesian Variable Selection Models for Order Restricted Problems

Chapter 2

Introduction to Order Restricted Bayesian Variable Selection

2.1 Model uncertainty in dose-response modelling

Dose-response experiments are an important part of a biomedical research to study relationships between increasing doses of a therapeutic compound and a variety of responses. Typically, the response represents a phenotypical effect of a compound such as inhibition, stimulation, toxicity, or expression level of a certain gene. The primary goal of such an experiment is to detect a dose-response relationship and to determine the nature of the relationship wherever it exists. In the following chapters, we focus on a continuous response and an experimental design with a fixed number of doses. We further assume that the dose-response relationship, if exists, is monotone, i.e. the compound effect (increasing or decreasing) becomes stronger (or stays the same) with an increasing dose (Kuiper *et al.*, 2014). Such property is very common in real applications, especially when inhibition or toxicity is measured. More general umbrella-shaped profiles (Bretz and Hothorn, 2003) can occur within a context of an over-dosing and therefore a decreasing (increasing) effect is expected after reaching some threshold dose. This setting will not be considered further in this chapter.

There are two main approaches for the analysis of dose-response experiments. The first approach uses parametric nonlinear models in order to estimate the dose-response relationship (Pinheiro *et al.*, 2006; Whitney and Ryan, 2009). The second approach assumes an underlying one-way analysis of variance (ANOVA) model with order restricted parameters (Robertson *et al.*, 1988; Bretz and Hothorn, 2003; Peddada *et al.*, 2005; Lin *et al.*, 2012b) and can be used in order to test the null hypothesis of no dose effect against an ordered alternative.

We consider the ANOVA setting in this chapter. The response is measured in $K-1$ dose levels and a control dose (placebo). Let μ_0 be the mean response under the control dose and $\mu_1, \mu_2, \ldots, \mu_{K-1}$ represent the mean responses under increasing doses of a therapeutic compound with $K-1$ dose levels. The primary interest is to detect a monotone dose effect. We call the case in which the therapeutical compound does not have any biological relevance to the response (e.g. a desired relationship for toxicity responses) as "no dose effect". Note that the no dose effect can appear for subset of doses only (e.g. some amount of compound is necessary to start the process or when all receptors become occupied and increasing the dose does not change the response). Therefore, for a given number of dose levels, the model space of an order restricted one-way ANOVA model consists of 2^{K-1} models defined by monotone constraints. For example, for the dose-response experiment with one control dose and three increasing dose levels (i.e. $K=4$), the model space is decomposed into 8 models presented in Table 2.1. The primary interest is to test the null hypothesis of no dose effect given by

$$H_0 : \mu_0 = \mu_1 = \mu_2 = \ldots = \mu_{K-1}, \tag{2.1}$$

against an ordered alternative

$$H^{up} : \mu_0 \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_{K-1} \quad \text{or} \quad H^{dn} : \mu_0 \geq \mu_1 \geq \mu_2 \geq \ldots \geq \mu_{K-1}, \tag{2.2}$$

with at least one strict inequality. Here, H^{up} and H^{dn} correspond to an upward and downward directions of the order constraints, respectively (Shkedy *et al.*, 2012b). Post hoc pairwise comparisons of means (e.g. Tukey's HSD, see Miller, 1981) lack power due to ignoring the monotonicity assumption. Instead, the likelihood-ratio test (LRT, Barlow *et al.*, 1972 and Robertson *et al.*, 1988) and multiple contrast tests (MCT, Mukerjee *et al.*, 1987, Bretz, 1999) are commonly used to test the null hypothesis of the no dose effect. However, the inference of these testing procedures ignores the model uncertainty since the best model among all the possible models is unknown. In fact, the inference for the LRT is based on one specific model from all the possible models under the alternative (the isotonic regression model that maximizes the likelihood under the order restrictions). The inference for the MCT takes into account different contrasts that correspond to different possible models under the alternative and the inference is based on one contrast only. Such a contrast can represent multiple models from set of possible models (Bretz and Hothorn, 2003). Both tests are post selection inference procedures that first select a model (or contrast) and then perform the inference.

Table 2.1: The set of eight possible monotonic dose-response models for an experiment with four dose levels (including placebo). Denote μ_i the mean response of dose level. The model g_0 represents the null model of no dose effect.

Model	Up: Mean Structure	Down: Mean Structure
g_0	$\mu_0 = \mu_1 = \mu_2 = \mu_3$	$\mu_0 = \mu_1 = \mu_2 = \mu_3$
g_1	$\mu_0 < \mu_1 = \mu_2 = \mu_3$	$\mu_0 > \mu_1 = \mu_2 = \mu_3$
g_2	$\mu_0 = \mu_1 < \mu_2 = \mu_3$	$\mu_0 = \mu_1 > \mu_2 = \mu_3$
g_3	$\mu_0 < \mu_1 < \mu_2 = \mu_3$	$\mu_0 > \mu_1 > \mu_2 = \mu_3$
g_4	$\mu_0 = \mu_1 = \mu_2 < \mu_3$	$\mu_0 = \mu_1 = \mu_2 > \mu_3$
g_5	$\mu_0 < \mu_1 = \mu_2 < \mu_3$	$\mu_0 > \mu_1 = \mu_2 > \mu_3$
g_6	$\mu_0 = \mu_1 < \mu_2 < \mu_3$	$\mu_0 = \mu_1 > \mu_2 > \mu_3$
g_7	$\mu_0 < \mu_1 < \mu_2 < \mu_3$	$\mu_0 > \mu_1 > \mu_2 > \mu_3$

Denote the whole set of models as $\mathcal{G}_R$. The problem of estimating dose-response profile is equivalent to the selection of monotone models that best describe the data given $\mathcal{G}_R$. When one particular model is selected and inference is done under the selected model, the uncertainty due to the model selection is ignored (Claeskens and Hjort, 2008). Such an approach can lead to bias in estimation of dose-specific means, especially when two models are almost equally supported by data.

Approaches that address the model uncertainty within the dose-response framework are discussed by Pinheiro *et al.* (2006), Bornkamp *et al.* (2009), Whitney and Ryan (2009) and Pinheiro *et al.* (2014). Bornkamp *et al.* (2009) use multiple comparison procedures to test candidate parametric models and base the estimates on weighted average of all suitable models (Raftery, 1995, Burnham and Anderson, 2002). Their approach is a synergy of parametric estimation and model selection frameworks. Generalization of the framework is introduced by Pinheiro *et al.* (2014). Whitney and Ryan (2009) focus on the estimation of a benchmark dose while taking into account the model uncertainty. They use an approximation of posterior probabilities of the model (Buckland *et al.*, 1997, Burnham and Anderson, 2002) based on the Bayesian Information Criterion (BIC, Schwarz, 1978), with non-informative priors for the set of $R+1$ candidate models, $g_0, \ldots, g_R$. This implies that prior probability of the model is set to $P(g_r) = 1/(R+1)$ for $r = 0, \ldots, R$. Specifically, the posterior probability of the model is given by $P(g_r|\text{data})$ and estimated by

$$\bar{P}(g_r|\text{data}) = \frac{\exp\left[-\frac{1}{2}\text{BIC}(g_r)\right] \cdot P(g_r)}{\sum_{k=0}^{R} \exp\left[-\frac{1}{2}\text{BIC}(g_k)\right] \cdot P(g_k)}. \tag{2.3}$$

Hereafter, we will refer to $P(g_r)$ as 'prior model probability' and to $P(g_r|\text{data})$ as 'posterior model probability'.

Pinheiro *et al.* (2006) focus on the estimation of the minimum effective dose. They proposed to estimate the mean response at each dose level by a weighted average $\bar{\mu} = \sum_{r=0}^{R} w_r \hat{\mu}_r$, where $\hat{\mu}_r$ are the estimates under model g_r and w_r are the posterior model probabilities given in (2.3). Similar methods for dose-response analysis for microarray data are discussed in Lin *et al.* (2012c) and Pramana *et al.* (2012b). In general, these methods are cumbersome due to necessity of a separate analysis for each model. Moreover, the non-linear modelling approaches rely on parametrical assumptions about the dose-response shape that does not have to apply in our framework. Such models can be difficult to fit when a number of observations is small. Furthermore, the methods focus mainly on the estimation, while we aim to address the inference as well, while taking the model uncertainty into account. As alternative, we propose a Bayesian variable selection method for an analysis of the dose-response experiments, the Bayesian approach to estimate $P(g_r|\text{data})$ instead of Equation (2.3).

The Bayesian variable selection (BVS) is a flexible modelling framework for dose-response data. It implicitly accounts for model uncertainty and has broad range of application areas (e.g. Clyde and George, 2004, Casella and Moreno, 2006, Kasim *et al.*, 2012, Otava *et al.*, 2014, Rockova *et al.*, 2012, Rockova and George, 2014). We apply the BVS within the dose-response modelling setting, where order restricted one-way ANOVA model is used to estimate the relationship between a continuous response and dose (Otava *et al.*, 2014). The BVS method performs simultaneous analyses of all the possible models, provides the parameter estimates based on model averaging and generates a model selection tools using the posterior probability of each model. The approach is closely related to Gibbs variable selection proposed by Whitney and Ryan (2009). However, in contrast with the Gibbs variable selection approach, the BVS approach estimates the posterior probability for each one of the models in $\mathcal{G}_R$. The posterior mean response at each dose level is a weighted average of the posterior means of all models, weights being the posterior model probabilities. In addition, the posterior probability of the null model is of the primary interest, since it also represents a probability for false positives, i.e. wrongly rejecting the null hypothesis, and therefore can be used for inference (Newton *et al.*, 2007).

The chapter is organized as follows. The current frequentist procedures are discussed in Section 2.2, formulation of the hierarchical Bayesian model for dose-response data in Section 2.3 and the Bayesian variable selection approach are discussed in Section 2.4. In Section 2.5, we present the results from the application of the methodology to the case studies. A simulation study for the comparison between BVS and the frequentist methods is introduced in Section 2.6 and the chapter is concluded with a discussion in Section 2.7.

2.2 Testing the null hypothesis against a simple ordered alternative

The basic setting we considered in this chapter consists of a response variable measured in a sequence of dose levels. Let Y_{ij} represents the response for jth observation at dose level i and μ_i denotes the mean response at dose level i. In order to model the relationship between the response and the increasing doses of a therapeutic compound we formulate the following linear model

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \ \ \varepsilon_{ij} \sim N(0, \tau^{-1}), \ \ i = 0, \ldots, K-1, \quad j = 1, 2, \ldots, n_i \tag{2.4}$$

For a given direction, the likelihood-ratio test (LRT) computes the maximum likelihood estimates for the mean response under the two hypotheses formulated in Equation (2.2). The maximum likelihood estimator computed under the null hypothesis H_0 equals the sample mean $\hat{\mu} = \left(\sum_{i=0}^{K-1} \sum_{j=1}^{n_i} Y_{ij}\right) / \sum_{i=0}^{K-1} n_i$. The maximum likelihood estimator under the order restricted alternative H^{up} is the vector of isotonic means (Robertson *et al.*, 1988). The likelihood-ratio test statistic, proposed by Barlow *et al.* (1972), can be expressed as

$$T_{LRT} = \frac{RSS_0 - RSS_1}{RSS_0} = 1 - \frac{RSS_1}{RSS_0}, \tag{2.5}$$

where RSS_0 represents the residual sum of squares under the null hypothesis and RSS_1 the residual sum of squares under the alternative H^{up} (or H^{dn}). The null hypothesis is rejected for a large value of T_{LRT}. The null distribution of T_{LRT} is a mixture of Beta distributions with mixture probabilities $P(\ell, K, \boldsymbol{w})$, $\ell = 1, \ldots, K$, that are also known as the level probabilities. They represent the probability (under the null hypothesis) that the number of unique isotonic means equals to ℓ in an experiment with K possible levels. According to Barlow *et al.* (1972), the p-value can be calculated by

$$P_{H_0}(T_{LRT} \geq t_{LRT}) = \sum_{\ell=1}^{K} P(\ell, K, \boldsymbol{w}) P\left[B_{\frac{1}{2}(\ell-1), \frac{1}{2}(N-\ell)} \geq t_{LRT}\right] \tag{2.6}$$

with N being the total number of observations, ℓ the number of final levels and $B_{\frac{1}{2}(\ell-1), \frac{1}{2}(N-\ell)}$ denotes a Beta distribution with $\alpha = 1/2(\ell-1)$ and $\beta = 1/2(N-\ell)$ and $B_{0,\beta} \equiv 0$. The inverse $\boldsymbol{w}^{-1} = (w_0^{-1}, ..., w_K^{-1})$ equals the variance of the response at each dose. For $K = 4$ and equal weights $\boldsymbol{w}_0$, the probability for one level only equals $P(\ell = 1, 4, \boldsymbol{w}_0) = 0.25$, $P(\ell = 2, 4, \boldsymbol{w}_0) = 0.46$, $P(\ell = 3, 4, \boldsymbol{w}_0) = 0.25$ and $P(\ell = 4, 4, \boldsymbol{w}_0) = 0.04$ (Robertson *et al.*, 1988). Note that level probabilities themselves are related to the isotonic regression results and not to the testing of the null hypothesis. They show the probability of obtaining certain number of the isotonic means under the null hypothesis. Hence, they do not depend on the variability of the data unless the variability differs across the doses. For more details about isotonic regression and level probabilities see Chapter 5.

A second approach to test the null hypothesis is multiple contrast test (MCT). The motivation for developing multiple contrast tests by Mukerjee *et al.* (1987) was to achieve tests with similar power to the LRT, but easier to use and interpret (Lin *et al.*, 2012b). The key idea is to perform as small number of comparisons as possible while covering sufficiently the alternative hypothesis. The test is based on simultaneous use of V single contrast tests (SCTs) defined as

$$T_v^{SC} = \frac{\sum_{i=0}^{K-1} c_i \hat{\mu}_i}{s \cdot \sqrt{\sum_{i=0}^{K-1} \frac{c_i^2}{n_i}}}, \tag{2.7}$$

where $v = 1, \ldots, V$, $\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$, $s = \sqrt{\frac{1}{\nu} \sum_{i=0}^{K-1} \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_i)^2}$ and $\nu = \sum_{i=0}^{K-1} (n_i - K)$.

The contrast vector $\boldsymbol{c} = (c_0, \ldots, c_{K-1})$ fulfills the condition $\sum_{i=0}^{K-1} c_i = 0$. Bretz (2006) shows that, under normality assumption, the test statistic T^{SC} follows an univariate central t-distribution with ν degrees of freedom under H_0. The MCT test statistic is the maximum over these V SCTs:

$$T^{MC} = \max_{v=1,\ldots,V} \{T_1^{SC}, T_2^{SC} \ldots T_V^{SC}\}. \tag{2.8}$$

Covering the space of the alternative hypotheses translates into a choice of a combination of vectors $\boldsymbol{c}_v$, $v = 1, \ldots, V$ (Lin *et al.*, 2012b). The MCT for the set of the single contacts tests $(T_1^{SC}, T_2^{SC} \ldots T_V^{SC})$ can be defined using a contrast matrix given by

$$\boldsymbol{C}^{MC} = \begin{pmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \\ \vdots \\ \boldsymbol{c}_V \end{pmatrix} = \begin{pmatrix} c_{10} & c_{11} & \ldots & c_{1,K-1} \\ c_{20} & c_{21} & \ldots & c_{2,K-1} \\ \vdots & & \vdots & \\ c_{V0} & c_{V1} & \ldots & c_{VK} \end{pmatrix}. \tag{2.9}$$

Each row of the contrast matrix $\boldsymbol{C}^{MC}$ corresponds to one contrast vector $\boldsymbol{c}$ of the SCT. The choice of the set of the vectors $\boldsymbol{c}_v$ determines properties of the test and distinguish between the different MCTs (Hothorn, 2006). For further comparison, we use two of them: Williams' and Marcus' MCTs (Bretz, 1999) based on the tests designed by Williams (1971) and Marcus (1976). Designs of the tests determine the choice of $\boldsymbol{c}_v$, $v = 1, \ldots, V$. Williams' MCT is based on the comparison between first (usually control) dose and the weighted average over the last b $(b = 1, ..., K-1)$ doses. It originates from a comparison of the last dose mean $\hat{\mu}_{K-1}^*$ computed using the isotonic regression, under the different possible profiles $g_1, \dots, g_R$, with an estimate of the mean of the first dose $\hat{\mu}_0$. Hence, due to the properties of the 'pool adjacent violators algorithm' (PAVA) it holds that

$$\hat{\mu}^*_{K-1} - \hat{\mu}_0 = \max \boldsymbol{C}^{Wil} \hat{\boldsymbol{\mu}}, \tag{2.10}$$

where $\hat{\boldsymbol{\mu}} = (\hat{\mu}_0, \dots \hat{\mu}_{K-1})^T$ and

$$\boldsymbol{C}^{Wil} = \begin{pmatrix} -1 & 0 & \dots & 0 & 1 \\ -1 & 0 & \dots & \frac{n_{K-2}}{n_{K-2}+n_{K-1}} & \frac{n_{K-1}}{n_{K-2}+n_{K-1}} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ -1 & \frac{n_1}{n_1+\dots+n_{K-1}} & \dots & \frac{n_{K-2}}{n_1+\dots+n_{K-1}} & \frac{n_{K-1}}{n_1+\dots+n_{K-1}} \end{pmatrix}. \tag{2.11}$$

The matrix $\boldsymbol{C}^{Wil}$ is called Williams-type MCT matrix and we use it to construct our set of the MCTs through Equation (2.10).

Marcus' MCT is a modification of Williams' idea with replacing the estimate of the mean of the first dose $\hat{\mu}_0$ with the isotonic estimate $\hat{\mu}^*_0$. Unfortunately, there is no general close form solution for $\boldsymbol{C}$ for Marcus' constraints, since its structure depends on the number of the doses. It can be easily constructed using each element of the following relationship as one contrast:

$$\hat{\mu}^*_{K-1} - \hat{\mu}^*_0 = \max\left\{0, \max_{0 \le g,h \le K-1}\left\{\frac{n_g\hat{\mu}_g + \dots + n_{K-1}\hat{\mu}_{K-1}}{n_g + \dots + n_{K-1}} - \frac{n_0\hat{\mu}_0 + \dots + n_h\hat{\mu}_h}{n_0 + \dots + n_h}\right\}\right\}. \tag{2.12}$$

The inference of Williams' and Marcus' MCTs can be based on the multivariate t-distribution. For the details about the distribution and about both procedures, we recommend to see Bretz (2006) or Lin *et al.* (2012b).

2.3 Bayesian estimation under strict inequality constraints

The aim in this section is to estimate the parameters under a strict inequality constraints $\mu_0 < \mu_1 < \mu_2 < \dots < \mu_{K-1}$. The constraints can be achieved by constraining the parameter space of $\boldsymbol{\mu} = (\mu_0, \dots, \mu_{K-1})$, whereby the order restrictions are imposed on the prior distributions. For a monotone upward profile we assume that for a profile function $\psi(i)$ it holds that $\psi(i) = \mu_{\lfloor i \rfloor}$ and that $\psi(i)$ is a right-continuous, nondecreasing function defined on interval $[0, K-1]$. We do not assume any deterministic relationship between μ_i and the dose levels, instead we specify a probabilistic model for μ_i at each distinct dose level.

To estimate $\boldsymbol{\mu}$ under the order restrictions, $\mu_0 < \mu_1 < \ldots < \mu_{K-1}$, the K dimensional parameter vector is constrained to lie in a subset $S^K \in \mathbb{R}^K$. The constrained set S^K is determined by the order among the components of $\boldsymbol{\mu}$. In this case, it is natural to incorporate the constraints into the specification of the prior distribution (Klugkist and Mulder, 2008). Let $\boldsymbol{Y} = (Y_{11}, Y_{12}, \ldots, Y_{K-1,n_{K-1}})$ be the response value and $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ the hyperparameters for $\boldsymbol{\mu}$. Gelfand *et al.* (1992) showed that the posterior distribution of $\boldsymbol{\mu}$, given the constraints, is the unconstrained posterior distribution normalized such that

$$P(\boldsymbol{\mu}|\boldsymbol{Y}) \propto \frac{P(\boldsymbol{Y}|\boldsymbol{\mu})P(\boldsymbol{\mu}|\boldsymbol{\eta},\boldsymbol{\tau})}{\int_{S^k} P(\boldsymbol{Y}|\boldsymbol{\mu})P(\boldsymbol{\mu}|\boldsymbol{\eta},\boldsymbol{\tau})d\boldsymbol{\mu}}, \quad \boldsymbol{\mu} \in S^K. \tag{2.13}$$

Let $S_i^K(\mu_l, l \neq i)$ be a cross section of S^K defined by the constraints for μ_i at a specified set of μ_l, with $l = 0, 1, 2, \ldots, i-1, i+1, \ldots, K-1$. In our setting, $S_i^K(\mu_l, l \neq i)$ is part of the interval $[\mu_{i-1}, \mu_{i+1}]$. It follows from Equation (2.13) that the posterior distribution for μ_i is given by

$$\begin{cases} P(\mu_i|\boldsymbol{Y},\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\mu}_{-i}) \propto P(\boldsymbol{Y}|\boldsymbol{\mu})P(\boldsymbol{\mu}|\boldsymbol{\eta},\boldsymbol{\tau}), & \mu_i \in S_i^K(\mu_l, l \neq i), \\ 0, & \mu_i \notin S_i^K(\mu_l, l \neq i). \end{cases} \tag{2.14}$$

where $\boldsymbol{\mu}_{-i} = (\mu_0, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_{K-1})$. Hence, when the likelihood and the prior distribution are combined, the posterior conditional distribution of $\mu_i|\boldsymbol{Y},\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\mu}_{-i}$ is the standard posterior distribution restricted to $S_i^K(\mu_l, l \neq i)$, i.e. restricted to the interval $[\mu_{i-1}, \mu_{i+1}]$ (Gelfand *et al.*, 1992). As a result, the sampling from the full conditional distribution can be reduced to the interval restricted sampling from the standard posterior distribution. Following Klugkist and Mulder (2008), we formulate an order restricted ANOVA model for which the mean response at the ith dose level is given by

$$E(Y_{ij}) = \mu_i = \begin{cases} \mu_0, & i = 0, \\ \mu_0 + \sum_{h=1}^{i} \theta_h, & i = 1, \ldots, K-1 \end{cases} \tag{2.15}$$

with the constraints that $\theta_h \geq 0$ for an upward trend or $\theta_h \leq 0$ for a downward trend. In a matrix notation, the mean gene expression for an upward trend model (for $K = 4$ and

$n = 3)$ is given by

$$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \begin{cases} \mu_0, & \text{control,} \\ \mu_0 + \theta_1, & \text{first dose level,} \\ \mu_0 + \theta_1 + \theta_2, & \text{second dose level,} \\ \mu_0 + \theta_1 + \theta_2 + \theta_3, & \text{third dose level.} \end{cases} \tag{2.16}$$

In order to complete the specification of the hierarchical model, we assume the following prior distribution for the unknown model parameters,

$$\begin{aligned} &\mu_0 \sim N(\eta_{\mu_0}, \tau_{\mu_0}^{-1}), \\ &\theta_h \sim TN(\eta_{\theta_h}, \tau_{\theta_h}^{-1}, 0, A), \quad h =, 1, \dots, K-1. \end{aligned} \tag{2.17}$$

Here $TN(\mu, \sigma^2, a, b)$ is a truncated normal distribution with mean μ, variance σ^2 and a, b the limits of the truncation interval. A is a positive constant. The model is fitted using a Markov Chain Monte Carlo (MCMC) simulation. The constant A is used to right truncate the distribution to achieve better properties of the MCMC chains. Its value is context dependent and has to be large enough not to influence the estimates. Practical way of selection A is to set it as difference between minimum and maximum of the data, since any reasonable estimate for any θ_h cannot exceed this number. The priors are further determined by hyperparameters with a non-informative specification. Normal distribution with large variance is used for the mean parameters, so the prior is as uniform as possible. Similar consequence has choice of Gamma distribution for the variance parameters.

$$\begin{aligned} &\tau \sim \Gamma(10^{-3}, 10^{-3}), \\ &\eta_{\mu_0} \sim N(0, 10^6), \\ &\tau_{\mu_0} \sim \Gamma(1, 1), \\ &\eta_{\theta_h} \sim N(0, 10^6), \quad h = 1, \dots, K-1 \\ &\tau_{\theta_h} \sim \Gamma(1, 1), \quad h = 1, \dots, K-1. \end{aligned} \tag{2.18}$$

2.4 Bayesian variable selection models for dose-response modelling

The Bayesian inequality model defined above cannot be used in our framework due to the equality constraints on the means of the null model and some of the alternative models. As pointed out by Dunson and Neelon (2003), since the priors of the components of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{K-1})$ are the truncated normal distributions, the mean structure $\mu_i = \mu_0 + \sum_{h=1}^{i} \theta_h$ implies an order constraints mean structure with the strict inequalities $\mu_0 < \mu_1 <, \ldots, < \mu_{K-1}$. The equality constraints would, in practice, assign zero probabilities to all other competing models except the model with the strict inequality constraints, model g_R (Klugkist and Hoijtink, 2007). In what follows we propose a Bayesian variable selection model that can be seen as an extension of the informative hypothesis inference framework discussed by Klugkist and Hoijtink (2007) to the setting in which equality constraints can be incorporated in the mean structure. Then, all the different models under the alternative hypothesis are taken into account for both inference and estimation. The equality constraints can be incorporated in the model by setting some of the components in $\boldsymbol{\theta}$ to be equal to zero. Indeed, $\theta_i = 0$ implies $\mu_i = \mu_{i-1}$.

The differences in the mean structures of the different models, therefore, depends on which of the components in $\boldsymbol{\theta}$ are set to be equal to zero or equivalently which columns in the ordered design matrix $\boldsymbol{X}$ are excluded. Hence, the design matrix $\boldsymbol{X}_{g_r}$ for the model g_r is in fact a subset of the design matrix $\boldsymbol{X}$. For example, for an experiment with $K = 4$ dose levels and $n = 3$ replicates, the design matrices for all the models presented in Table 2.1 are given, respectively, by

$$\boldsymbol{X}_{(g_0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{X}_{(g_1)} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{X}_{(g_2)} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix},$$

$$\boldsymbol{X}_{(g_3)} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \ \boldsymbol{X}_{(g_4)} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \ \boldsymbol{X}_{(g_5)} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

$$\boldsymbol{X}_{(g_6)} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \ \boldsymbol{X}_{(g_7)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

The mean gene expression for each model g_r is given by

$$E(Y_{ij}|g_r) = \boldsymbol{X}_{g_r}\boldsymbol{\beta}'_r, \quad r = 0, \ldots, R,$$

where $\boldsymbol{\beta}_r$ is the parameter vector for each model given by

$$\boldsymbol{\beta}_r' = \begin{cases} \mu_0, & \text{model } g_0, \\ (\mu_0, \theta_1)', & \text{model } g_1, \\ (\mu_0, \theta_2)', & \text{model } g_2, \\ (\mu_0, \theta_1, \theta_2)', & \text{model } g_3, \\ (\mu_0, \theta_3)', & \text{model } g_4, \\ (\mu_0, \theta_1, \theta_3)', & \text{model } g_5, \\ (\mu_0, \theta_2, \theta_3)', & \text{model } g_6, \\ (\mu_0, \theta_1, \theta_2, \theta_3)' & \text{model } g_7. \end{cases}$$

As a result, the problem of the model estimation in the presence of equality constraints is reduced to a problem of variable selection depending on which of the columns of $\boldsymbol{X}$ are selected or deleted. This is related to the Bayesian variable selection approach (George and McCulloch, 1993) which is used to determine an optimal model from a priori set of $R+1$ known plausible models. As pointed out by O'Hara and Sillanpää (2009) the choice of an optimal model reduces to the choice of a subset of variables which are included in the model (i.e. model selection), or the choice of which parameters in the parameter vector are different from zero (i.e. inference). This can be done by rewriting the mean structure in (2.15), using δ_h and z_h instead of θ_h (O'Hara and Sillanpää, 2009, Ohlssen and Racine, 2015, Otava *et al.*, 2014), as

$$E(Y_{ij}) = \mu_0 + \sum_{h=1}^{i} \theta_h = \mu_0 + \sum_{h=1}^{i} z_h \delta_h. \tag{2.19}$$

where z_h, $h = 1, \ldots, K-1$, is an indicator variable such that

$$z_h = \begin{cases} 1, & \delta_h \text{ is included in the model,} \\ 0, & \delta_h \text{ is not included in the model.} \end{cases} \tag{2.20}$$

For the four dose level experiment ($K = 4$) discussed above, the triplet $\boldsymbol{z} = (z_1, z_2, z_3)$ defines uniquely each one of the eight plausible models. For example, for $\tilde{\boldsymbol{z}}_1 = (0,0,0)$ holds that $E(Y_{ij}|\mathcal{G}_R, \boldsymbol{z} = \tilde{\boldsymbol{z}}_1) = (\mu_0, \mu_0, \mu_0, \mu_0)$ (which corresponds to the mean of the model g_0) and for $\tilde{\boldsymbol{z}}_2 = (1,0,0)$ we obtain $E(Y_{ij}|\mathcal{G}_R, \boldsymbol{z} = \tilde{\boldsymbol{z}}_2) = (\mu_0, \mu_0+\delta_1, \mu_0+\delta_1, \mu_0+\delta_1)$ (which corresponds to the mean of the model g_2). Hence, in our setting the BVS model estimates the posterior probability of each model, $P(g_r|\text{data})$, and in particular the posterior probability of the null model, $P(g_0|\text{data})$. For example, $P[\boldsymbol{z} = (0,0,0)|\text{data}] = P[E(Y_{ij}) = \mu_0|\text{data}]$.

Kuo and Mallick (1998) approach was used for the specification of the prior models for z_h and δ_h. It assumes that z_h and δ_h are independent, i.e. $P(\delta_h, z_h) = P(\delta_h) \times P(z_h)$, with a truncated normal prior distribution for δ_h, same as for θ_h in Equation (2.17). In case of lack of any prior information about the models probability, non-informative priors can be used for z_h. Following Jeffreys (1961) (as discussed by Kass and Wasserman, 1996), we recommend to use equal weights for all the models. The prior specification is defined as:

$$\begin{aligned} &z_h \sim \text{Bernoulli}(\pi_h), \\ &\pi_h \sim \text{U}(0,1). \end{aligned} \tag{2.21}$$

The variable π_h represents inclusion probability of z_h and can be estimated by the proportion of the $z_h = 1$ within the Markov Chain Monte Carlo (MCMC) simulation run.

As pointed out by O'Hara and Sillanpää (2009), the posterior inclusion probability of δ_h in the model is the posterior mean of z_h. Further, for a given value of K, using the indicator variables z_h, we specify a transformation function that uniquely defines each one of the plausible models (Ntzoufras, 2002), $G = 1 + \sum_{h=1}^{K-1} z_h 2^{h-1}$. Thus, the posterior probability of $G = r + 1$ defines uniquely the posterior probability of a specific model g_r (when g_r defined as in Table 2.1). In particular (for K=4), the posterior probability of the null model is given by

$$\bar{P}(G = 1|\text{data}) = \bar{P}\left[E(Y_{ij}) = \mu_0|\text{data}\right] = \bar{P}\left[\boldsymbol{z} = c(0,0,0)|\text{data}\right] = \bar{P}(g_0|\text{data}). \tag{2.22}$$

Note that we omitted in Equation (2.22) the dependency on the models and we write $\bar{P}(G = 1|\text{data})$ instead of $\bar{P}(G = 1|\text{data}, g_0, \ldots, g_R)$. This simplification of notation will be used for the remainder of the thesis.

For $K = 4$ there are eight possible monotone models (for a given direction): seven monotone models (given in Table 2.1) and the null model. It follows that G is given by

$$G = \begin{cases} 1, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 0, z_3 = 0), \quad \text{model } g_0, \\ 2, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 0, z_3 = 0), \quad \text{model } g_1, \\ 3, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 1, z_3 = 0), \quad \text{model } g_2, \\ 4, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 1, z_3 = 0), \quad \text{model } g_3, \\ 5, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 0, z_3 = 1), \quad \text{model } g_4, \\ 6, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 0, z_3 = 1), \quad \text{model } g_5, \\ 7, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 1, z_3 = 1), \quad \text{model } g_6, \\ 8, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 1, z_3 = 1), \quad \text{model } g_7. \end{cases} \tag{2.23}$$

Note that the estimation of mean vector $\boldsymbol{\mu}$ is computed as its posterior mean $\bar{\boldsymbol{\mu}}$ of B MCMC simulations. It holds that $\bar{\boldsymbol{\mu}} = \frac{1}{B}\sum_{b=1}^{B} \hat{\boldsymbol{\mu}}_b$, while in each iteration b, one model g_r is considered and estimate $\hat{\boldsymbol{\mu}}_b$ is obtained. The model g_r is selected n_{g_r} times over all the B iterations, with estimate $\hat{\boldsymbol{\mu}}_{g_r}$. Therefore $\bar{\boldsymbol{\mu}} = \frac{1}{B}\sum_{r=0}^{R} n_{g_r}\hat{\boldsymbol{\mu}}_{g_r}$. Since posterior probability $\bar{P}(g_r|\text{data}) = n_{g_r}/B$, i.e. it corresponds to proportion of selection of the model, the equation can be rewritten as $\bar{\boldsymbol{\mu}} = \sum_{r=0}^{R} \bar{P}(g_r|\text{data})\hat{\boldsymbol{\mu}}_{g_r}$. Therefore, mean estimates $\bar{\boldsymbol{\mu}}$ are in fact model averaging based estimates, weighted by the posterior probabilities of the models.

In summary, the BVS model provides a simultaneous framework for the estimation and the model selection. The estimates at each dose level are represented by the posterior means that are in fact a weighted Bayesian model average of all the plausible models. The weights equal to the proportion of visits of particular model during the MCMC simulation, i.e. the posterior model probability of the model g_r is estimated as $\bar{P}(G = r+1|\text{data}, g_0, \ldots, g_R)$.

2.5 Application to the case studies

Three real life studies are used to illustrate the methodology discussed in this chapter. All the case studies have the same data structure: response is measured under increasing doses of the respective compounds with the first dose being a control (placebo). The data sets are presented in Section 1.1 and Section 1.2. The data set from each study was analyzed using the LRT, the MCT with Williams' and Marcus' contrast and the BVS model. The BVS models were fitted in `Winbugs 1.4` (Lunn *et al.*, 2000) using MCMC simulation with 20,000 iterations from which the first 5,000 were discarded as burn-in period.

2.5.1 The Ames data

The results obtained for all the methods are presented in Table 2.2. All the frequentist methods show an evidence against the null hypothesis. The posterior probability of the null model obtained for the BVS model ($6.7 \cdot 10^{-5}$) indicates no evidence in favour of the null model, but substantive evidence in support of an alternative model with monotone relationship between the frequency of mutation and the increasing doses of the compound (0.408).

Figure 2.1a reveals a close agreement between the posterior means obtained for the BVS model and maximum likelihood parameter estimates obtained by the isotonic regression for the Ames study. Note that the posterior means obtained from the BVS model do not correspond to the one specific model but it is the Bayesian weighted model averaging of all competing models (for $K = 5$ there are 16 possible models, including the null model). Interestingly, similar to the isotonic regression which pools together the means of the last three dose levels, the inclusion probabilities (Figure 2.1b) obtained from the BVS

Table 2.2: P-values for the frequentist methods and the posterior model probabilities for the BVS model. "BVS null" shows the posterior probability of the null model and "BVS max" shows the maximal posterior probability among the posterior probabilities of all the alternative monotone models.

	LRT	MCT(W)	MCT(M)	BVS null	BVS max
Ames	$6 \cdot 10^{-5}$	$1.4 \cdot 10^{-5}$	$3.6 \cdot 10^{-5}$	$6.7 \cdot 10^{-5}$	0.408
Litter	0.029	0.019	0.029	0.220	0.623

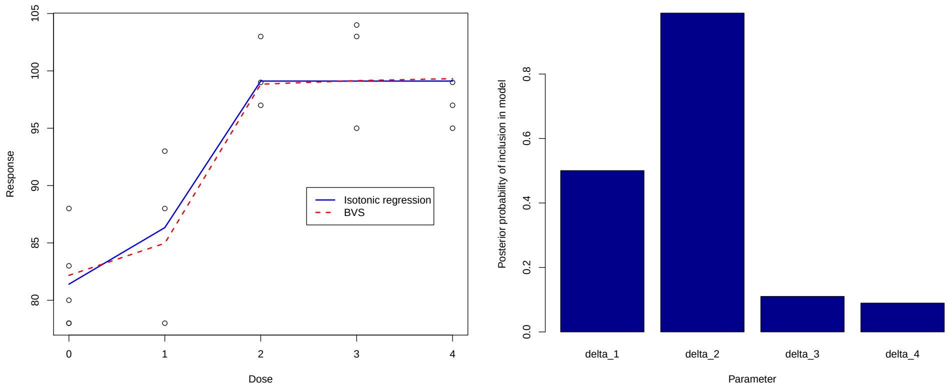


Figure 2.1: *The Ames mutagenity data. Left panel: Observed data, isotonic regression (solid line) and posterior mean of the BVS model (dashed line). Right panel: Posterior mean of z_h, i.e. the inclusion probability of δ_h into the model.*

model show little evidence in support of different dose effects for dose 3 and dose 4 (with the estimated posterior probabilities of 0.11 and 0.09, respectively). Therefore, models with increments between first two doses, g_1 and g_2 have highest posterior probability (see Figure 3.1d).

2.5.2 The Litter data

The p-values and the posterior model probabilities for the Litter data are shown in 2.2. The LRT and MCTs reject the null hypothesis. The posterior probability of the null hypothesis obtained from BVS is 0.22, which implies that there is more support in favor of the alternative hypothesis given the data. Specifically, the BVS shows more substantive evidence in support of the alternative model g_1 (defined in Table 2.1) whose posterior

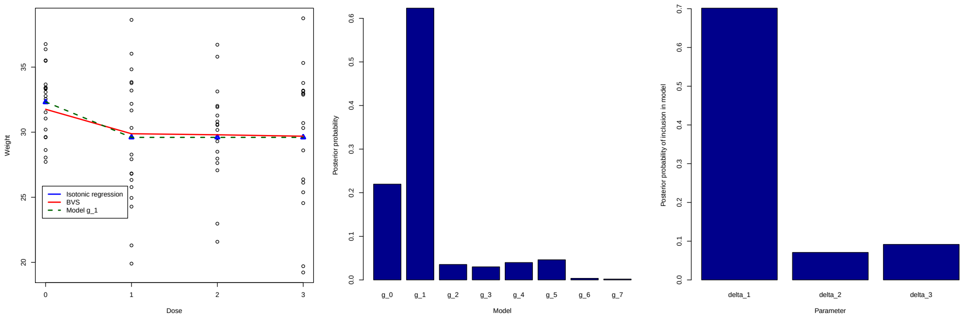


Figure 2.2: *The Litter data. Left panel: Observed data, isotonic regression (solid line) and posterior mean of the BVS model (dashed line). Dotted line; the posterior mean obtained by MCMC when only model g_1, i.e. model with maximum posterior probability for BVS, was taken into account. Dotted line coincides with solid line almost perfectly. Middle panel: Posterior probability of null model g_0 and alternative models g_r, $r = 1, \ldots, 7$. Notation corresponds to the model numbers presented in Table 2.1. Right panel: Posterior mean of z_h, i.e. the inclusion probability of δ_h into the model.*

model probability is 0.623 (see Figure 2.2b). This model has a common dose effects for dose 1 to dose 3. This illustrates an important aspect of the BVS model which simultaneously performs the inference and provides the evidence for all the possible models given the data. Furthermore, the inclusion probabilities, shown in Figure 2.2c, indicate that the δ_2 and δ_3 should not be included in the model which corresponds to the results obtained from the isotonic regression.

Due to the fact that the posterior probability of model g_1 is relatively high compared to the other models, the posterior means of the BVS model are similar to those of the isotonic means and the posterior means from g_1 with the common mean for dose 1 to dose 3 and the different mean for control (Figure 2.2a). Note that model g_1 is different from the BVS model since its design matrix is fixed while the BVS fits all the possible models simultaneously and produce the model averaging of the posteriors means for doses across all the competing models, weighted by their respective posterior model probabilities, given the data.

2.5.3 The direct posterior probability approach for multiplicity adjustment

The aim of the analysis of the HESCA data set (see Section 1.2) is to detect genes with monotone expression profiles. Due to a high dimensionality of microarray data, the dose-response microarray analysis of the HESCA study requires multiplicity adjustment both within and between genes. Typically, the family wise error rate (FWER) that represents the overall Type I error, i.e. the probability of at least one false rejection of the null hypothesis, and the false discovery rate (FDR), i.e. the expected proportion of the false rejections among all the rejections, are used for the multiplicity adjustment. Following Lin *et al.* (2012b) we apply the FWER method for the multiplicity adjustment within the genes and the FDR for the multiplicity adjustment between the genes. In the following section, we discuss the use of the posterior probability of the null model for the FDR adjustment within the BVS framework.

Assume that there are $m = 1, \ldots, M$ genes to analyze simultaneously and the aim is to identify genes that exhibit a monotone relationship with increasing doses of a therapeutic compound. The problem is equivalent to investigating if expression levels of each gene show substantive evidence against the null model g_0. The posterior probability of the null model $P_m(g_0|\text{data})$ holds dual properties as the likelihood of the null model and simultaneously the probability of the false rejection of the null hypothesis, i.e. when there is no dose-response relationship, but the gene is identified as following the monotone profile (Newton *et al.*, 2007). For a pre-specified threshold α, $P_m(g_0|\text{data})$ represents probability of the false positive for the gene m. Let I_m be an indicator variable for $P_m(g_0|\text{data}) \leq \alpha$ (i.e. indicator for including gene m among genes with "significant" dose-response relationship). The expected number of the false discoveries (cFD) is defined as

$$\text{cFD}(\alpha) = \text{E}(\text{cFD}) = \sum_{m=1}^{M} P_m(g_0|\text{data}) I_m. \tag{2.24}$$

Newton *et al.* (2007) define the conditional (on the data) false discovery rate as

$$\text{cFDR}(\alpha) = \frac{\text{cFD}(\alpha)}{N(\alpha)}, \tag{2.25}$$

where $N(\alpha)$ is the number of genes declared significant for a given threshold α. Then, the cFDR(α) represents an expected error done by using the threshold α to identify significant genes. Then, the cFDR(α) represents a mean error made by considering any gene as significant using the threshold α. Hence, we select a value of α in a way to keep the cFDR(α) under the pre-specified threshold ω.

Figure 2.3 shows the relationship between the false discovery rate, the number of significant genes and the threshold for the HESCA case study. As expected, the higher the threshold, the higher the cFDR and the number of significant genes. The implication of this relationship is that in order to control for a certain level of the cFDR, the corresponding threshold can be used as significance level for the posterior probability under the null

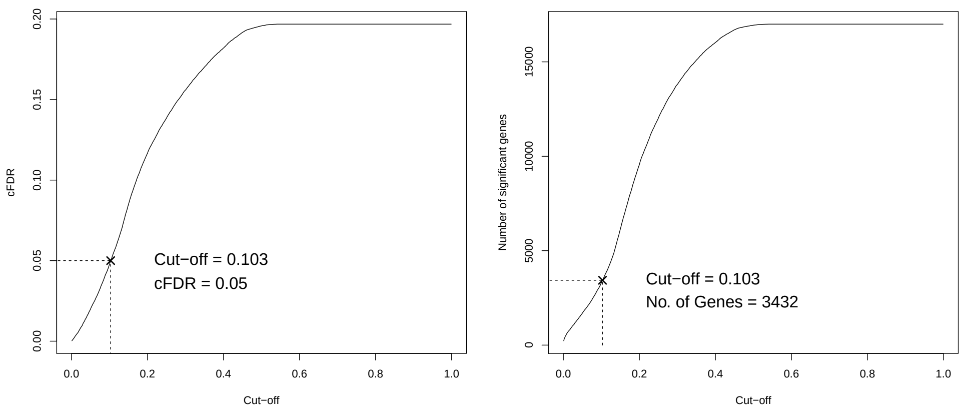


Figure 2.3: *Adjustment for the multiplicity for the HESCA data. Left panel: The relationship between the conditional false discovery rate (cFDR) and the cut-off values. Right panel: The relationship between the number of significant genes and the cut-off values.*

model. Note that the concept of significance level used here is data dependent and consequently, the cFDR control is conditional on the data.

Table 2.3 shows the number of genes with significant dose-relationships under the upward and downward monotone profiles from the HESCA study. The number of genes with significant dose-response relationships is higher for the frequentist methods than the BVS model at 5% false discovery rate. At a fixed level of FDR, the higher number of the significant genes for the frequentist methods may imply better power with these methods than the BVS model since the power is often associated with the number of significant genes. However, the FDR controlled by the frequentist methods and the cFDR introduced for the BVS context are not entirely the same quantities, since one arise from an adjustment of p-values and the other one from an adjustment of the posterior probabilities of the null model. The comparison of the BVS model and the frequentists methods in terms of Type I error and power is investigated through a simulation study presented in the next section.

2.6 Simulation study

A simulation study was conducted in order to investigate the performance of the BVS model in terms of Type I error and power. The data were generated according to the order restricted one-way ANOVA model specified in Equation (2.4), $Y_{ij} \sim N(\lambda\mu_i, \tau^{-1})$, with $\tau = 1$ and varying $\lambda = 1, 2, 3$. The parameter λ is used to control the magnitude

Table 2.3: Number of rejected null hypotheses according to the frequentist methods and to the BVS model while controlling the FDR and the cFDR on a 0.05 level, respectively.

Profile	LRT	MCT(W)	MCT(M)	BVS
Upward	2057	1772	1954	1634
Downward	2464	2053	2364	1798

of increment of the mean response from one dose level to another. The higher the value of λ, the higher the increment. The null hypothesis is formed as the equality of means $\mu_0 = \cdots = \mu_{K-1}$. The simulation settings corresponds to models shown in Table 2.1 and they are described in details in Section 6.1. For each setting, 1,000 data sets were simulated. An experiment with $K = 4, 5$ dose levels and $n = 3, 4, 5$ observations per dose was investigated. For $K = 4$ and $n = 4$, simulation was repeated with different choices of variance of Gaussian distribution. In this chapter, we discuss in detail mainly the results for the case of $K = 4$ and $n = 3$. All the remaining results are shown in Section 6.2.

The BVS model, one-sided LRT and one-sided MCTs were performed. Table 2.4 shows the empirical Type I error obtained for each method. All the methods control Type I error at 5%, while the MCTs are more conservative than the LRT. The BVS model seems even more conservative. To achieve similar proportion of false rejections as in the case of the LRT and the MCTs, i.e. 0.05, we can use a threshold as high as 0.35 for the BVS rejection (see Table 2.4 and Figure 2.4).

Table 2.5 shows the power of the methods for $K = 4$. As expected the LRT seems to be the most powerful test with both MCTs slightly worse and BVS with threshold 0.05 is about 0.10 behind the MCTs. The parameter λ represents increasing magnitude of dose-response effect (see Chapter 6 for details). With an increasing λ, the difference between the methods diminishes (this pattern is visualized in Figure 2.5 for $n = 4$). Such result is expected, because with higher λ, the power approaches one for all the methods. The improving performance of the BVS with an increasing threshold is natural, too. If higher threshold is used, we achieve results comparable in terms of power with frequentist methods (Figure 2.5), while still controlling Type I error at a pre-specified value. Similar result was obtained for the cases of $K = 5$ and $n = 5$ (for details, see Chapter 6). Figure 2.6 demonstrates visually the change in the power when the number of dose levels increase from $K = 4$ to $K = 5$ which corresponds to a change from $1/8$ to $1/16$ for the model prior probabilities, respectively. Note that the first seven models corresponds to $K = 4$ (circles) while the last 15 models corresponds to $K = 5$ (filled circles). We can see that a change in the power across the models and dose levels for the BVS model behaves

Table 2.4: Type I error of the frequentist methods and the BVS model for $K = 4$ and $K = 5$. Four BVS columns correspond to the choice of threshold used for rejection of H_0.

	LRT	MCT(W)	MCT(M)	BVS 0.05	BVS 0.10	BVS 0.15	BVS 0.35
$K = 4$							
$n = 3$	0.041	0.037	0.036	0.002	0.003	0.003	0.034
$n = 4$	0.044	0.044	0.048	0.002	0.002	0.005	0.027
$n = 5$	0.053	0.057	0.051	0.001	0.001	0.001	0.017
$K = 5$							
$n = 3$	0.047	0.048	0.048	0.000	0.002	0.005	0.046
$n = 4$	0.056	0.051	0.052	0.000	0.002	0.003	0.030
$n = 5$	0.048	0.056	0.051	0.000	0.002	0.004	0.022

in a very similar way as the change in the power obtained for the LRT. Hence, the power is influenced by additional information provided by the data (by adding more dose levels) and not only by the change in prior probability of g_0 (from $1/8$ to $1/16$). Similar patterns were observed for different values of λ and n (e.g. Figure 6.10, Figure 6.11).

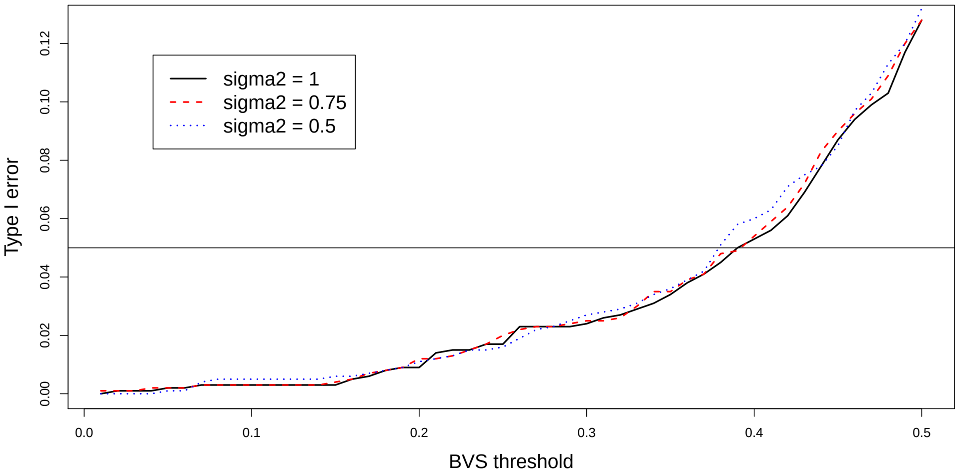


Figure 2.4: *Proportion of the false rejections (Type I error) for the BVS when varying the threshold of a rejection of the null model. Based on* $1,000$ *simulated data sets with* $n = 4$ *observations per dose with three values of* $\tau^{-1} = \sigma^2$*:* $\sigma^2 = 1$ *(solid line),* $\sigma^2 = 0.75$ *(dashed line) and* $\sigma^2 = 0.50$ *(dotted line).*

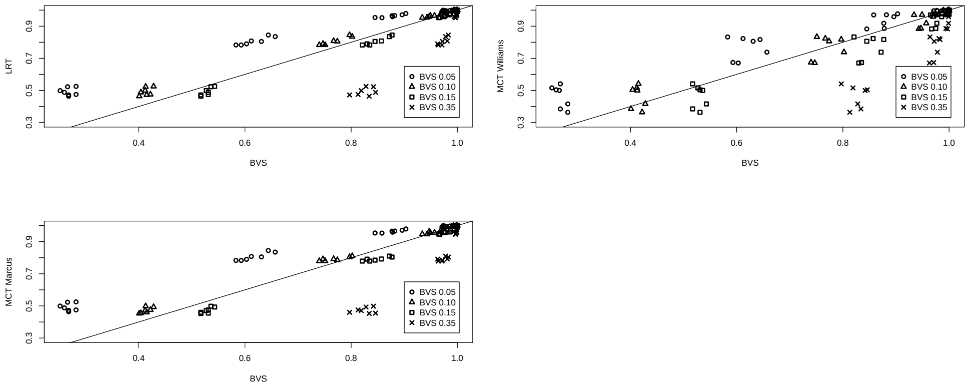


Figure 2.5: *Comparison of the power between the BVS (with varying threshold) and the frequentist tests. Circles represent the results for threshold* $\alpha = 0.05$*, triangles* $\alpha = 0.10$ *and rectangles* $\alpha = 0.15$*. The plot is based on simulation under* $n = 4$*. Top left: LRT vs. BVS. Top right: MCT Williams vs. BVS, Bottom left: MCT Marcus vs. BVS.*

Table 2.5: Results for $K = 4$ and $n = 3$. The columns RT and MCTs show estimation of the power of the particular tests. The columns BVS shows proportion of posterior probabilities of the null model given the data that are smaller then $\alpha = 0.05, 0.10, 0.15, 0.35$.

			MCT	MCT	BVS	BVS	BVS	BVS
λ	Profile	LRT	(W)	(M)	0.05	0.10	0.15	0.35
1	g_1	0.36	0.42	0.34	0.22	0.37	0.49	0.81
	g_2	0.38	0.31	0.36	0.22	0.37	0.48	0.80
	g_3	0.40	0.39	0.35	0.22	0.38	0.50	0.83
	g_4	0.36	0.26	0.35	0.22	0.38	0.50	0.83
	g_5	0.44	0.42	0.39	0.26	0.41	0.54	0.86
	g_6	0.41	0.33	0.38	0.22	0.36	0.49	0.82
	g_7	0.46	0.42	0.41	0.24	0.40	0.52	0.85
2	g_1	0.85	0.90	0.85	0.74	0.88	0.93	0.99
	g_2	0.86	0.73	0.84	0.74	0.88	0.94	0.99
	g_3	0.89	0.88	0.86	0.81	0.90	0.95	0.99
	g_4	0.85	0.72	0.82	0.74	0.85	0.92	0.99
	g_5	0.90	0.91	0.87	0.80	0.92	0.96	1.00
	g_6	0.90	0.81	0.87	0.80	0.91	0.96	1.00
	g_7	0.90	0.88	0.87	0.82	0.93	0.97	1.00
3	g_1	0.99	1.00	0.99	0.98	0.99	1.00	1.00
	g_2	0.99	0.98	0.99	0.99	1.00	1.00	1.00
	g_3	0.99	0.99	0.99	0.98	0.99	1.00	1.00
	g_4	0.99	0.97	0.99	0.97	0.99	1.00	1.00
	g_5	1.00	1.00	0.99	0.98	1.00	1.00	1.00
	g_6	1.00	0.98	0.99	0.99	1.00	1.00	1.00
	g_7	1.00	0.99	0.99	0.99	1.00	1.00	1.00

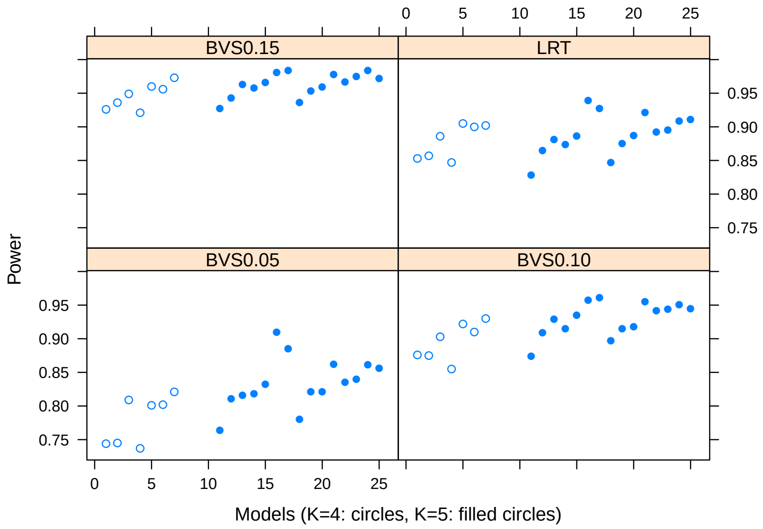


Figure 2.6: *Comparison of the power between $K = 4$ and $K = 5$ for BVS (with varying threshold) and LRT (top right panel). The plot is based on a simulation under $\lambda = 2$ and $n = 3$. The models are ordered arbitrarily, seven models for $K = 4$ on the left (circles) and 15 models for $K = 5$ on the right (filled circles).*

2.7 Discussion

In many applications, an analysis of the dose-response data requires to test the null hypothesis of no dose effect against an ordered alternative and to estimate the dose-response curve. In this chapter, we focus on a Bayesian approach for an order constrained one-way ANOVA models. The inequality constraints were incorporated as priors in the Bayesian formulation of the model. We have shown that the approach of (Gelfand *et al.*, 1992) assigns zero probabilities to the models with equality constraints, so it is not suitable in our setting. In order to overcome the problem of the zero probabilities for the equality constraints, we introduced the BVS model formulation for the dose-response modelling.

The BVS model as presented assumes an independent prior model for the joint distribution of z_i and δ_i, i.e. $P(\delta_i, z_i) = P(\delta_i) \times P(z_i)$ and non-informative priors for both $P(\delta_i)$ and $P(z_i)$. An alternative approach is to formulate a model for $P(z_i, \delta_i)$ by taking into account the conditional distribution $P(\delta_i|z_i)$. Dellaportas *et al.* (2002) proposed the Gibbs Variable Selection (GVS) method which assumes a mixture model for the conditional distribution $P(\delta_i|z_i)$, i.e. $P(\delta_i|z_i) = z_i N(\eta_i, S) + (1-z_i)N(0, \tau^{-1})$. The Stochastic Search Variable Selection (SSVS) by (George and McCulloch, 1993) assumes the following mixture model for $P(\delta_i|z_i) = z_i N(0, \tau^{-1}) + (1-z_i)N(0, g\tau^{-1})$. In both GVS and SSVS is necessary to specify priors for the tuning parameters (S and τ for GVS, g and τ for SSVS). In both cases a prior knowledge about the increment is needed for specification. The influence of the choice of the prior model for $P(\delta_i|z_i)$ will not be investigated further in this thesis.

We have shown that using the BVS model allows us to calculate the posterior probability for each one of the candidate models and in particular the posterior probability of the null model. Therefore, the BVS model proposed in this chapter can be used for both inference and estimation of dose-response curve. Further, the posterior mean obtained from the BVS model is a model average of all the candidate order restricted one-way ANOVA models for a given value of the dose levels.

The simulation study showed that the BVS can match the frequentist methods in terms of power while controlling similar level of Type I error. The comparison is valid, because we avoid to compare p-values and the posterior probabilities themselves, but rather the results based on using any of these two quantities for answering a question of the null hypothesis testing. The power of the BVS method indeed depends on chosen threshold. The approach on how to avoid the necessity of threshold specification, while keeping good operational characteristics, is introduced in Chapter 3.

The suggestion on how to automatize threshold specification for microarray data by controlling conditional FDR was described in Section 2.5.3. However, control of the cFDR does not imply control of the FDR. This property arises from the fact that significant genes tend to have $P(g_0|\text{data})$ nearly zero, allowing to enter large amount of non-significant genes with $P(g_0|\text{data})$ around 0.6 (simulation not shown). Therefore, we do not focus on this method further in this thesis and develop instead methodology described in Chapter 3.

Additionally, the BVS provides an evidence for the possible models under the monotone constraints. The probability of identification of the true monotone profile based on the posterior probabilities may bring further insight into the BVS model properties. Together with the comparison between the posterior probability of the most likely model and the posterior probability of other models, the BVS may also be used for model selection among the alternative monotone models. This topic is further discussed in Chapter 4.

The presented BVS model was based on the use of non-informative priors for the selection variables $z_1, \ldots, z_K$. Strong scientific knowledge is typically rare in dose-response modelling situations, but when it is present (e.g. if historical data are available), it can be very easily incorporated. Adjustment of the hyperprior for π_i or prior for z_i translates into change in the prior probabilities of the models. Indeed, such a change can highly influence the posterior probability of the different models and so the estimated dose-specific means (since they are in fact weighted average of model-specific means with weights equaled to the posterior probability of the models). Hence, we suggest to use informative priors only in cases, when scientific knowledge is really strong and to specify them very carefully. Analysis of the effect of priors on posteriors in case of non-informative priors is investigated in Chapter 5.

The proposed model and the analysis framework focus on normally distributed response. Generalization in spirit of Pinheiro *et al.* (2014) for binary data, count data, longitudinal data or clustered outputs can be achieved due to flexibility of the Bayesian framework. The analysis workflow would stay the same, only the model specification and the prior distributions on the mean structure would need to be modified. Similarly, the order restriction assumption can be modified by varying truncation of the priors on the mean structures.

Chapter 3

Inference for Bayesian Variable Selection

3.1 Introduction

In this chapter, we focus on the inference procedures based on the posterior probability $P(g_0|\text{data}, g_0, \ldots, g_R)$ of the null model. In what follows, we show that $P(g_0|\text{data}, g_0, \ldots, g_R)$ equals the posterior probability of the null hypothesis $P(H_0$ is correct$|\text{data}, g_0, \ldots, g_R)$. Given an estimate $\bar{P}(g_0|\text{data}, g_0, \ldots, g_R)$, we wish to choose a threshold ω, so that $\bar{P}(g_0|\text{data}, g_0, \ldots, g_R) < \omega$ implies a rejection of the null hypothesis. Instead of focusing on the choice of the threshold itself that could lead to rather arbitrary decisions, as shown in previous chapter (and by Otava *et al.*, 2014), in this chapter we focus on the distribution of the posterior probability of the null model, $P(g_0|\text{data}, g_0, \ldots, g_R)$, under the null hypothesis. We introduce a permutation based inference procedure that is objective in the sense that it is robust to a choice of configuration of priors of the models $g_0, \ldots, g_R$. Hence, we are able to obtain a measure that quantifies the evidence contained by the posterior probability that is not influenced by a non-informative prior distribution specification. The procedure is based on permutation tests and it is introduced in Section 3.2. The proposed BVS model and the inference procedure are applied to the case studies in Section 3.3. A simulation study conducted to assess the performance of the proposed method is presented in Section 3.4. Finally, we discuss further properties of the method, advanced topics and possible future extensions in Section 3.5.

3.2 Methodology

3.2.1 Inference for BVS model

Our main interest is to test the null hypothesis of the no dose effect given in Equation (2.1) against the ordered alternative. The quantity we propose to use as a test statistic is the posterior probability of the null model $\bar{P}(g_0|\text{data})$, given in Equation (2.22). We first discuss the rejection rule in Section 3.2.1.1. In Section 3.2.1.2 follows a discussion about a permutation method that can be used in order to approximate the distribution of $P(g_0|\text{data})$ under the null hypothesis.

3.2.1.1 Inference based on $P(g_0|\text{data})$

Bayesian inference for the null model has been based on a fixed threshold ω, such that the null hypothesis is rejected whenever $\bar{P}(g_0|\text{data}) < \omega$ (Do *et al.*, 2006, Goldstein, 2006). However, the choice of the appropriate threshold remains debatable. It is unclear how to choose the value of ω in order to maintain a desirable level of Type I error and power.

To demonstrate the problem, let us focus on the results obtained for the Litter data presented in Figure 3.1c. The posterior probability for g_0 is estimated as 0.217. What should be inference decision based on this value is, without a prior knowledge, very unclear. Should we reject the null hypothesis since $\bar{P}(g_0|\text{data}) < \bar{P}(g_1|\text{data})$ or $\bar{P}(g_0|\text{data}) = 0.217$ provides enough evidence in favour of the null hypothesis? Moreover, even if we choose the value of ω, how can we choose this value in such a way that we control the Type I error at a pre-specified level?

Otava *et al.* (2014) showed via simulations that rejection rules based on the posterior probability can control the Type I error rate with higher thresholds than would be corresponding frequentist choice ($\omega = 0.35$ controls the same Type I error rate as the frequentist significance level, $\alpha = 0.05$, while achieving higher power). Of course, for an analysis of a real life data one does not know which value of ω to choose and therefore Type I error cannot be controlled in practice. Hence, for inference, our main focus is not just $\bar{P}(g_0|\text{data})$, but the distribution of $P(g_0|\text{data})$ under the null hypothesis as well.

3.2.1.2 Permutation test based

The proposed method compares the estimated value of the posterior probability $\bar{P}(g_0|\text{data})$ with the distribution of posterior probability of the null model under the null hypothesis, $P(g_0|H_0)$. We propose to estimate the distribution of $P(g_0|H_0)$ using a permutation procedure based on permutations of the doses. The permuted data are denoted as data^*. Specifically, we test how extreme is the value of the observed posterior probability in comparison to the posterior probabilities of the null model obtained from the permuted data sets. The underlying principle for this approach is that if the null hypothesis holds (i.e. no dose effect), the permutation of the doses and its associated probability under the null model $\bar{P}(g_0|\text{data}^*)$ simulates drawing from the null distribution.

The above permutation test is also referred to as an exact test (Fisher, 1936), because it evaluates all the possible permutations of the responses with respect to the dose levels. However, such an approach is computationally intensive, e.g. for $K = 4$ and $n = 3$ there are 369,000 possible permutations. Therefore, a random sampling of a fixed number of permutations B from the permutation space is usually chosen to approximate an exact distribution of the statistics (Dwass, 1957). Comprehensive summary of properties of the permutation test can be found in Ernst (2004).

Once the null distribution of $P(g_0|\text{data})$ is estimated, it will be compared with an observed value of $\bar{P}(g_0|\text{data})$ and a permutation p-value (p_{Bayes}) of the test for H_0 against H_1^{up} or H_1^{dn} will be computed. The permutation p-value p_{Bayes} is robust against the choice of the specification of the non-informative prior distribution, which is often a desirable property in the dose-response analysis due to lack of strong prior believe about the dose effects (we elaborate on the robustness and non-informative priors in Chapter 5). The complete resampling based inference algorithm follows:

1. Permute the observed response vector $\boldsymbol{Y}$ B times to get $\boldsymbol{Y}^{(1)}, \dots, \boldsymbol{Y}^{(B)}$.

2. For each permuted data $\boldsymbol{Y}^{(b)}$, fit the BVS model (with the same prior distributions).

3. Estimate the posterior probability of the null model, $\bar{P}[g_0|\boldsymbol{Y}^{(1)}] \dots, \bar{P}[g_0|\boldsymbol{Y}^{(B)}]$.

4. Denote $Q = \sum_{b=1}^{B} I\left\{\bar{P}[g_0|\boldsymbol{Y}^{(b)}] < \bar{P}[g_0|\boldsymbol{Y}]\right\}$, which measures how many times permuted BVS results suggest stronger evidence against H_0 than BVS results based on observed data.

5. Calculate the p-value by
$$p_{\text{Bayes}} = \frac{Q}{B}. \tag{3.1}$$

6. Reject H_0 if $p_{\text{Bayes}} < \alpha$.

Note that $p_{\text{Bayes}} = 0$ has to be interpreted in the context of B, a number of the permutations. It should be correctly stated as $p_{\text{Bayes}} < 1/B$, because it translates to zero events among the B permutations. Using p_{Bayes} and α has substantial advantages compared to using $\bar{P}(g_0|\text{data})$ and ω. The quantity p_{Bayes} is robust with respect to the choice of priors and it induces control of frequentist operating characteristics.

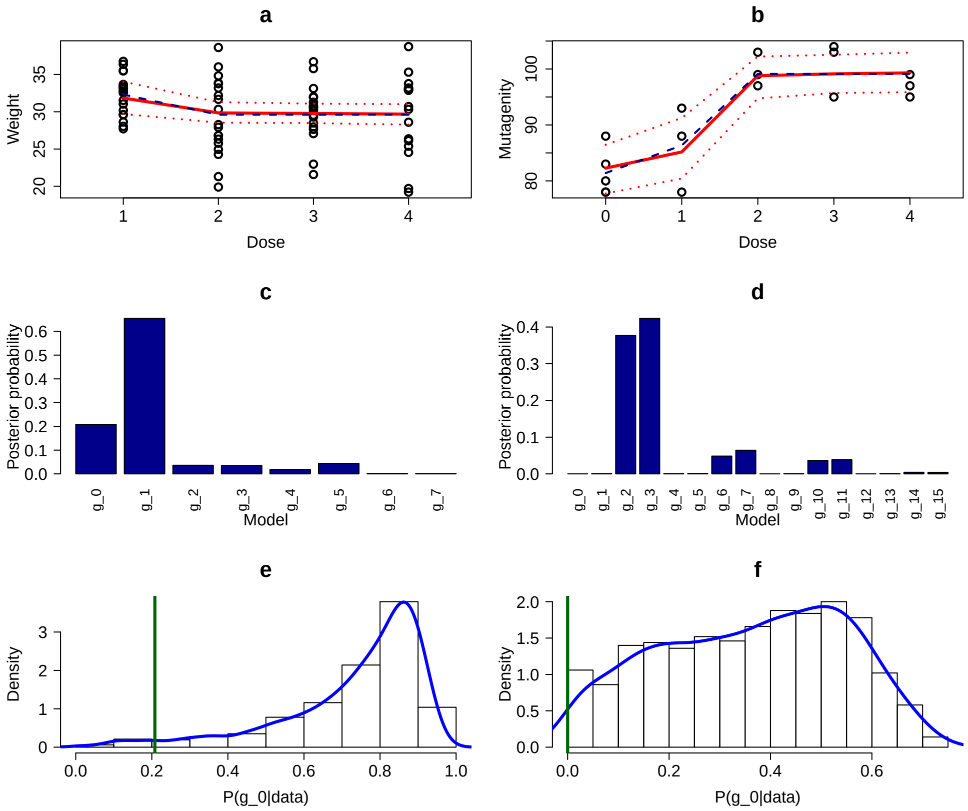


Figure 3.1: *The Litter data (left column, panel a, c, e) and the Ames data (right column, panels b, d, f). Panels a, b show the observed data and the posterior means of the BVS model (red solid line) with 95% credible intervals (dotted lines) and the isotonic regression estimate of the means (blue dashed line). Panels c, d show the posterior probability of the null model* g_0 *and the alternative models* g_r, $r = 1, \dots, 7$ *(*$r = 1, \dots, 15$*, respectively). Notation corresponds to the model numbers presented in Table 2.1. Panels e, f show the distribution of the posterior probabilities estimated by the permutation test with 1,000 permutations. Histogram is supported with a smooth density estimate (blue solid line) and the estimate of posterior probability* $\bar{P}(g_0|\mathrm{data})$ *is shown (green vertical solid line).*

3.3 Results

We applied the BVS methodology on the four data sets introduced in Section 1.1: Litter, Ames, Toxicity and Angina data sets. The analysis was performed using the package `runjags` (Denwood, In Review) of `R` software (R Core Team, 2013) together with the `JAGS` software (Plummer, 2003). We performed the permutation test as discussed in Section 3.2.1.2, with $B = 1,000$ permutations. The results of the BVS model were compared with the MCT (Bretz, 2006), based on Marcus' (Marcus, 1976) and Williams' (Williams, 1971) contrasts, and with the LRT. The p-values of these methods were denoted as p_{Marcus}, p_{Williams} and p_{LRT}, respectively.

The results of the BVS method for the Litter data are shown in the left panels of Figure 3.1 (panels a, c and e). Panel 3.1a shows the estimated dose-specific means based on the BVS (red) compared to the isotonic regression estimates (blue). Although they are very close to each other, they do not completely coincide. The isotonic regression is based on one particular model only (model g_1 here), while the BVS estimates are weighted averages of all the possible models. The weights are defined as the posterior probabilities of the respective models (panel 3.1c). The estimated posterior probability of the null model $\bar{P}(g_0|\text{data})$ is equal to 0.217 and it is greater than or equal to the estimated posterior probabilities of the null model $\bar{P}_b(g_0|\text{data}^*)$ in 2.5% of the $B = 1,000$ permuted data sets, resulting in $p_{\text{Bayes}} = 0.025$. This is equal to the area to the left of a vertical line in panel 3.1e. As the result, we reject the null hypothesis at the level of significance 5%. We conclude that decrease in the litter weight is associated with increasing dose. Since the null model of no dose effect is rejected, the estimate of the dose-response relationship between the litter weight and the dose is based on the model average weighted by the posterior probabilities of each model. Model g_1 is particularly dominant in this specific example. The result is comparable to the results obtained for the LRT and the MCTs: $p_{\text{LRT}} = 0.028$, $p_{\text{Marcus}} = 0.029$, $p_{\text{Williams}} = 0.018$. The BVS method has an added advantage of a unified analytical framework for the inference for the null model under the model uncertainty, the model selection among the alternative models when the null model is rejected and the model averaging of the estimated dose-response relationship across all the possible monotone models.

The results for the Ames data set are shown in the right panels of Figure 3.1 (panels b, d and f). The difference between the BVS model averaging based estimates and the isotonic regression is slightly more pronounced than in the Litter data (panel 3.1b). This is due to the fact that there is not one single dominating alternative model. In fact, there are two models with almost equal posterior probabilities, the model g_2 with $\bar{P}(g_2|\text{data}) = 0.42$ and g_3 with $\bar{P}(g_3|\text{data}) = 0.39$ (panel 3.1d). The observed posterior probability of the null model g_0 obtained for the the Ames data is equal to $3 \cdot 10^{-5}$ (Figure 3.1d) with $p_{\text{Bayes}} = 0$ (panel 3.1f). Such a results is expected, because the distribution of the posterior probability $P(g_0|\text{data}^*)$ from the permuted data sets would be indeed rather far from zero, i.e. observed $\bar{P}(g_0|\text{data})$. The result corresponds to the frequentist p-values $p_{\text{LRT}} = 3 \cdot 10^{-5}$, $p_{\text{Marcus}} = 2.4 \cdot 10^{-5}$, $p_{\text{Williams}} = 3.3 \cdot 10^{-5}$. We conclude a rejection of the null model and consequently that an increase in the mutagenicity is associated with the increasing dose. Higher dose is more likely to cause changes in genetic information than lower dose.

The results obtained for the BVS model for Angina data, estimates and posterior probabilities of particular models, are shown in Figure 3.2a, 3.2c and 3.2e. The posterior probability of the null model $\bar{P}(g_0|\text{data}, g_0, \dots, g_8)$ is equal to 0 (see Figure 2c). The BVS permutation test was applied with $B = 1,000$ permutations. Equation (3.1) gives $p_{\text{Bayes}} = 0$. The result is in agreement with the results obtained for LRT and MCTs: $p_{\text{LRT}} = 3 \cdot 10^{-5}$, $p_{\text{Marcus}} = 7.5 \cdot 10^{-9}$, $p_{\text{Williams}} = 1 \cdot 10^{-8}$. The estimate of the empirical distribution of the posterior probability of the null model g_0 under the null hypothesis is shown in Figure 3.2e.

For the Toxicity data set, $\bar{P}(g_0|\text{data}) = 0.122$. The $p_{\text{Bayes}} = 0.013$ which corresponds to frequentist p-values $p_{\text{LRT}} = 0.013$, $p_{\text{Marcus}} = 0.026$, $p_{\text{Williams}} = 0.016$. The visual representation of results is presented in Figure 3.2b, 3.2d and 3.2f.

Note that $\bar{P}(g_0|\text{data}) = 0$ is caused by the restriction of the length of MCMC chain used. As mentioned above, 20,000 iterations were used to compute posterior probability of null model and none of them selected the null model. Hence, correct statement about the $\bar{P}(g_0|\text{data})$ is that $\bar{P}(g_0|\text{data}) < \frac{1}{20{,}000}$, i.e. $5 \cdot 10^{-5}$. We typically do not address this issue and keep $\bar{P}(g_0|\text{data}) = 0$, because we are usually not interested in this level of precision. If more accurate results are necessary, longer chain should be used.

In summary, results of all the case studies support the conclusion that p_{Bayes} behaves in similar way as frequentist p-value, closely related to LRT test.

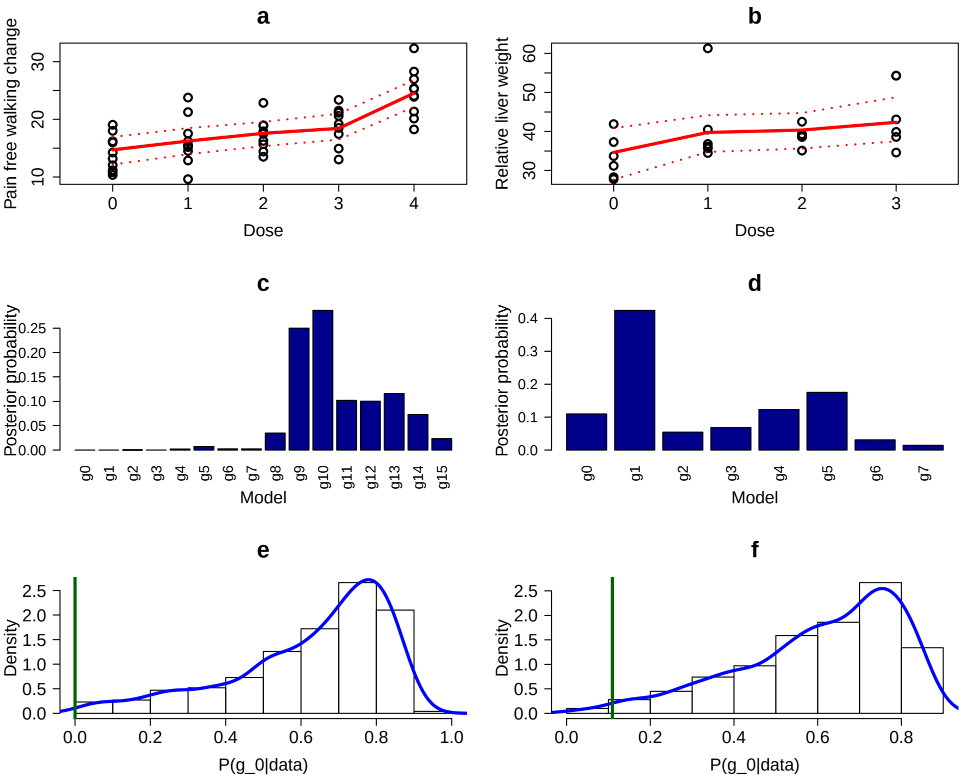


Figure 3.2: *The Angina data (left column, panel a, c, e) and Toxicity data (right column, panel b, d, f). Panels a, b show observed data and posterior means of the BVS model (red solid line) with 95% credible intervals (dotted lines). Panels c, d show posterior probability of null model* g_0 *and alternative models* g_r, $r = 1, \ldots, 7$ *(*$r = 1, \ldots, 15$*, respectively). Notation corresponds to the model numbers presented in Table 2.1 and Table 6.2. Panels e, f show distribution of posterior probabilities estimated by permutation test with 1,000 permutations. Histogram is supported with smooth density estimate (blue solid line) and estimate of posterior probability* $\bar{P}(g_0|\mathrm{data})$ *is shown (green vertical solid line).*

3.4 Simulation study

A simulation study was conducted in order to investigate the performance of the permutation based inference for the BVS method in terms of the Type I error rate and the power. The data were generated according to the order restricted one-way ANOVA model specified in Equation (2.4), $Y_{ij} \sim N(\lambda\mu_i, \tau^{-1})$, with $\tau = 1$ and varying λ. The parameter λ is used to control the magnitude of increment of the mean response from one dose level to another. The higher the value of λ, the higher the increment. The simulation represented an experiment with $K = 4$ and $K = 5$ dose levels and followed the design described in Section 6.1. The number of observations per dose level was equal to $n = 3$ and $n = 4$. The permutation test, introduced in Section 3.2.1.2, was performed using $B = 1,000$ permutations. The null hypothesis was rejected whenever $p_{\text{Bayes}} < \alpha$, with $\alpha = 0.05$. The performance of the BVS model was compared with the Williams' and Marcus' contrast based MCT and with the LRT.

Table 3.1 presents the simulation results for $n = 3$ and $K = 4$, Figure 3.3 displays the result for $\lambda = 2$. We see that for all the methods the empirical Type I error rate is slightly above 0.05. The larger simulation study that is presented in Section 6.3 shows that the Type I error rate is well controlled by all the methods. The results of the power analysis suggest a desirable behaviour of permutation method. The permutation test is comparable with the LRT test, in general the most powerful method among the frequentist tests. The results of the remaining simulations (i.e. for $n = 4$, $K = 5$) were consistent with the results presented in this section and they are discussed in detail in Section 6.3.

Table 3.1: Results of simulation study for $n = 3$, $K = 4$. First row shows the Type I error rate. Each following row shows the power to reject the null hypothesis, if data were generated under the particular profile and λ value. MCT were applied with Williams' (W) and Marcus' (M) contrast.

λ	Profile	MCT (W)	MCT (M)	LRT	BVS
	g_0	0.051	0.056	0.060	0.059
1	g_1	0.436	0.350	0.376	0.365
	g_2	0.317	0.379	0.395	0.404
	g_3	0.414	0.385	0.433	0.440
	g_4	0.300	0.371	0.379	0.389
	g_5	0.429	0.406	0.439	0.449
	g_6	0.345	0.403	0.438	0.447
	g_7	0.413	0.411	0.465	0.479
2	g_1	0.922	0.865	0.867	0.866
	g_2	0.782	0.856	0.856	0.867
	g_3	0.904	0.871	0.893	0.903
	g_4	0.758	0.846	0.869	0.866
	g_5	0.907	0.882	0.896	0.902
	g_6	0.821	0.864	0.888	0.899
	g_7	0.899	0.889	0.911	0.920
3	g_1	0.998	0.993	0.991	0.991
	g_2	0.972	0.986	0.988	0.989
	g_3	0.994	0.992	0.993	0.993
	g_4	0.968	0.991	0.990	0.989
	g_5	0.996	0.991	0.995	0.995
	g_6	0.977	0.987	0.993	0.993
	g_7	0.992	0.993	0.994	0.995

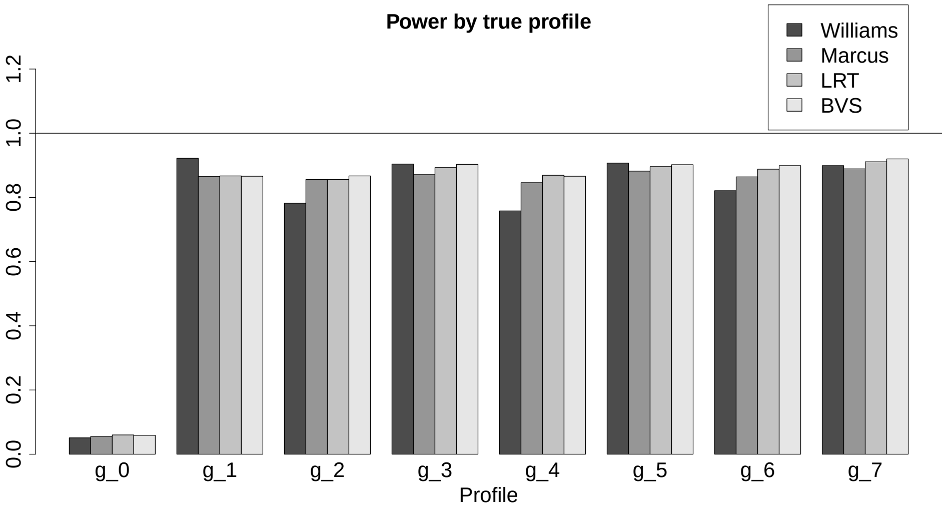


Figure 3.3: *Results of the simulation study for $n = 3$ and $K = 4$, with $\lambda = 2$. Each set of bars shows the power of rejecting the null hypothesis, if data were generated under the particular profile $g_1, \ldots, g_7$. In case of g_0, displayed quantity is the Type I error rate. Grey scale distinguishes among different tests: darkest for Williams' MCT, then Marcus' MCT, the LRT test and brigtest for the permutation test.*

3.5 Discussion

The main difference between the BVS method, proposed in this chapter, and the frequentist methods such as the LRT or the MCTs is the way of dealing with the model uncertainty. The LRT and the MCTs focus mainly on providing information about the rejection, not about the particular profiles. Moreover, only several profiles are actually taken into account while the test is performed (actually only one model for LRT). In contrast, the permutation test is using the BVS posterior probability, a quantity that is estimated while all the possible models are taken into account. This feature is very important, because our framework assumes that the zero effects of the increased dose are meaningful and therefore there are several candidate models to be considered. The LRT test cannot address this type of the model uncertainty, because it is only based on the model that the maximizes likelihood under the order restrictions and ignores all the other models. The MCTs represent compromise, because information about the contrast leading to the rejection can be obtained. However, since more profiles can be related to one contrast, the MCTs does not adjust for all the profiles as the BVS does.

Naturally, computing hundreds of permutations is more computationally intensive than the LRT or the MCTs. The computational burden is the main drawback of the proposed method. The time necessary for computation is partly dependent on the length of MCMC chain when fitting the BVS model. Our experience suggests good convergence properties across various settings and sample sizes already with about 20,000 iterations (after 5,000 iteration of burn-in period). For details, see Section 6.1.1. If necessary, the length of chain can be shortened (or prolonged) for a particular data set. Computation burden of the method also depends on the minimal value of p_{Bayes} that we can achieve. Such a value is simply an inverse of number of iterations for the permutation test. Note that this problem is embarrassingly parallel, so the computational time of the permutation test can be reduced using a parallel programming.

The inference for the dose-response data, as proposed in this chapter, should be robust towards prior misspecification when the priors for the null model are not specified close to zero or one. The setting the prior of the null model as zero (or one) would lead to $\bar{P}(g_0|\text{data}) = \bar{P}_b(g_0|\text{data})$ and so $p_{\text{Bayes}} = 0$ by definition and has nothing to do with the evidence in the data. However, we aim to use the method in case of lack of the prior knowledge. Situation of zero prior on the null hypothesis or the alternative hypothesis clearly cannot be considered as the non-informative case. In the case of common methods to establish the objective priors (equal priors, Jeffreys' priors, Kass and Wasserman, 1996), the robustness of p_{Bayes} should be retained. The exact quantification of the prior dependency is be pursued further in Chapter 5.

Chapter 4

Selection of the Minimum Effective Dose Based on the Posterior Probabilities

4.1 Introduction

The selection of the minimum effective dose (MED) is an important concept in the drug development process (European Medicines Agency, 2002 and Wang *et al.*, 2011). It translates into the identification of the lowest dose that causes a desired effect or adverse events. The MED is often used in the context of the former case, while the latter is called the lowest observed adverse event level (LOAEL, Kodell, 2009) or the maximum safe dose (Hothorn and Hauschke, 2000). From a statistical point of view, there is no difference between these two concepts, only the interpretation of the response and the findings differ. An analogous framework arises when the determination of the maximum effective dose is of primary interest (Kong *et al.*, 2014). In this chapter, we restrict the discussion to the MED. In some cases, the clinical significance is included in the definition of the MED (Liu, 2010), while other cases are focused on statistical significance only (Kuiper *et al.*, 2014). Note that clinical significance of the result can be included in stages following the analysis and treated separately.

The concept of the MED appears in multiple stages of drug development. If a large number of doses is used or prior knowledge about the shape of the dose-response profile exists, parametric methods can be applied (e.g. the four parameter logistic non-linear regression model, Hill's model, etc., Seber and Wild, 1989, Straetemans, 2012, Pramana *et al.*, 2012b). The MED is, in this case, based on a particular parametric model. Alternatively, methods can be used that combine model selection with parametric modelling, such as MCP-Mod (Bornkamp *et al.*, 2009). In our framework, there are only few dose levels in which the response was measured and typically only limited knowledge about the dose-response relationship exists. Therefore, parametric modelling of the whole profile as a continuous function of dose is not suitable and an order restricted analysis of variance (ANOVA) is preferred. Typically, the monotonicity assumption is a reasonable choice, implying that a higher dose induces a stronger effect (positive or negative for upward or downward trend, respectively). Note that this assumption is often made in drug development studies (e.g. Bretz and Hothorn, 2003 or Ohlssen and Racine, 2015).

The goal of the analysis is to determine the lowest active dose with significant difference to a control. For example, in an experiment with a placebo and three active doses, we would like to detect which of the three active doses is the MED. To achieve it, we need to be able to determine the probability of being the best model among the eight possible models (for each direction) shown in Table 2.1.

Within the frequentist framework, the MED can be viewed either in terms of inference of particular increments between consecutive doses or as model selection problem. The former approach is represented by multiple comparison procedures (Bretz and Hothorn, 2003), such as Dunnett's test (Dunnett, 1955). This approach may require to pool together some of the means in order to maintain a reasonable power, which does not provide complete information about the MED and can eventually lead to biased estimates (Hothorn and Hauschke, 2000). Multiple contrast tests are generally designed to preform an inference rather than to determine the MED (Bretz and Hothorn, 2003). Closed tests procedures can be applied instead, but they may lack overall power (Wang and Peng, 2015). Recently, Kuiper *et al.* (2014) suggested to focus on model selection methods and specifically on information criteria (IC) based approaches (e.g. Lin *et al.*, 2009, Lin *et al.*, 2012c). Within the IC approach, the posterior probability of each one of the candidate models is calculated and used for the determination of the MED. It is crucial to realize that the MED cannot be established through a classical model selection process that focuses only on the best model (among a set of candidate models). The competing models can have the same MED, i.e. the first dose showing significant effect compared to the mean of control dose (e.g. the MED for models g_1, g_3, g_5 and g_7 in Table 2.1 is the first active dose, see Kuiper *et al.*, 2014). Although a certain model can have the highest posterior probability, it could be worse than posterior probabilities of all the models with same MED pooled together. This reasoning suggests that IC is an appropriate approach, since IC based methods compare all candidate models and their IC values can be easily converted into weights that can be pooled together for appropriate models (Kuiper *et al.*, 2014). Naturally, order restriction needs to be taken into account for IC based methods (Anraku, 1999) which leads to the generalized order restricted information criterion (GORIC, Kuiper *et al.*, 2014). The advantage of the IC is that they provide the probability for a particular model being the best model, given the data, among all fitted models. Hence, multiple values of the MED can be computed together with their corresponding posterior probabilities (Kuiper *et al.*, 2014). The main disadvantage of this approach is that it requires to fit all the models under consideration. This is feasible in an experiment with relatively small number of dose levels, but it becomes infeasible for an experiment with relatively large number of dose levels. For example, for an experiment with five or six dose levels, there are 16 or 32 order restricted one-way ANOVA models that need to be fitted, respectively. Procedures are available to reduce the number of models either by an efficient search in the model space (e.g. stepwise methods) or by reducing the model space itself (e.g. diversity index, Kim *et al.*, 2014). However, they usually require additional input parameters or criteria specification and the resulting amount of models to be fitted can still remain prohibitive. In such a case, Bayesian variable selection method (George and McCulloch, 1993, O'Hara and Sillanpää, 2009) becomes an attractive alternative. In particular, for dose response experiments, the BVS approach (Kasim *et al.*, 2012, Otava *et al.*, 2014) allows fitting all models simultaneously and provides posterior probabilities for each of them, while computational time does not increase in a linear fashion as in case of the IC approach.

This chapter continues as follows. The methodological background for both the IC based methods and the BVS is summarized in Section 4.2. The methods are applied for the two case studies in Section 4.3 and the results are evaluated. Further empirical comparison is investigated via simulation study and presented in Section 4.4. Finally, the findings are summarized and discussed in Section 4.5.

4.2 Methodology

We consider a dose-response experiment with a control group and $K-1$ active dose levels. Denote the set of observations by

$$\boldsymbol{Y} = \{Y_{ij},\ \ i = 0,\dots,K-1,\ \ j = 1,\dots,n_i\},$$

where n_i represents the number of observations of dose i. Our goal is to select the lowest dose i that shows a statistically significant difference compared to the control group. Such a dose is the MED. We denote such an event as $\mathrm{MED} = i$ and the probability that this event occurs as $P(\mathrm{MED} = i)$. Let $g_0,\dots,g_R$ be a set of $R+1$

candidate models which are used to determine the MED. Based on the observed data and the models that are considered as plausible, the quantity of interest is the posterior probability of the particular value of the MED, $P(\text{MED} = i|\text{data}, g_0, \ldots, g_R)$. The determination of the MED can be translated into a model selection problem. For example, for $K = 4$ it translates to a selection of the best model among all models for given direction that are presented in Table 2.1. Note that multiple models induce the same MED, e.g. for $K = 4$ the probability that the MED is the second dose level is equal to $P(\text{MED} = 2|\text{data}, g_0, \ldots, g_R) = P(g_2|\text{data}, g_0, \ldots, g_R) + P(g_6|\text{data}, g_0, \ldots, g_R)$, where $P(g_r|\text{data}, g_0, \ldots, g_R)$ is the posterior probability of the model g_r, $r = 0, 1, \ldots, R$. Therefore, the inference about the MED cannot be based on a single model only and our aim is to estimate $P(g_r|\text{data}, g_0, \ldots, g_R)$ for all the suitable models. The posterior probabilities for the MED is obtained by summing appropriate posterior model probabilities. To simplify notation, from this point onwards, we denote $P(\text{MED} = i|\text{data}, g_0, \ldots, g_R)$ and $P(g_r|\text{data}, g_0, \ldots, g_R)$ as $P(\text{MED} = i|\text{data})$ and $P(g_r|\text{data})$, respectively.

4.2.1 Model averaging techniques

The likelihood based methodology addresses the problem of model selection through information criteria (IC) approaches (e.g. Akaike, 1974, Burnham and Anderson, 2002, Claeskens and Hjort, 2008, Lin *et al.*, 2012c, Kuiper *et al.*, 2014). All candidate models are fitted and their corresponding IC values are computed. Based on the IC value, weights are calculated for each of the fitted models (as explained in detail below). The resulting weights can be considered as an approximation of posterior probabilities of the models being the best model, among all fitted models given the data (Burnham and Anderson, 2002). Additionally, this approach enables us to incorporate prior knowledge if there is any available.

As proposed by Burnham and Anderson (2002) and Claeskens and Hjort (2008), for set of models $g_0, g_1, \ldots, g_R$, we can select as the best model such that maximizes the posterior model probability given by

$$P(g_r|\text{data}) = \frac{P(\text{data}|g_r)P(g_r)}{\sum_{s=1}^{R} P(\text{data}|g_s)P(g_s)} \qquad r = 0, \ldots, R. \tag{4.1}$$

The term $P(\text{data}|g_r)$ is the model likelihood (Burnham and Anderson, 2002) corrected with a penalization term and $P(g_r)$ is a prespecified prior probability of model g_r. In this section, we consider a vague prior knowledge and so we use $P(g_r) = 1/(R+1)$ for all r. The model likelihood $P(\text{data}|g_r)$ is approximated by

$$P_{IC}(\text{data}|g_r) = \exp(-\tfrac{1}{2}\Delta IC_r), \tag{4.2}$$

where $\Delta IC_r = IC_r - IC_{\min}$, with $IC_{\min} = \min_{r=0,\ldots,R} IC_r$. Hence, combining equations (4.1) and (4.2) together and assuming equal prior probabilities, we get

$$w_r = P_{IC}(g_r|\text{data}) = \frac{\exp(-\frac{1}{2}\Delta IC_r)}{\sum_{s=0}^{R}\exp(-\frac{1}{2}\Delta IC_s)}. \tag{4.3}$$

The properties of this method depends on IC used.

An information criterion is a function of likelihood with a penalization term for model complexity given by

$$IC = -2\text{log}L(\boldsymbol{\theta}|\text{data}) + \tau. \tag{4.4}$$

Here, $\boldsymbol{\theta}$ represents the model parameters and τ is a penalization function. IC such as Akaike's information criterion (AIC, Akaike, 1974) or Bayesian information criterion (BIC, Schwarz, 1978) can be applied. The AIC uses the penalty term $\tau = 2 \cdot A$, with A being number of parameters in a model. The main criticism against the AIC is that it evaluates the goodness of fit without taking into account sample size (Burnham and Anderson, 2004). Small-sample size modification of the criterion was developed (Sugiura, 1978), but often the original version is used (Burnham and Anderson, 2004). The BIC uses the penalty term $\tau = A \cdot \log(B)$, where B is the number of observations. Hence, the BIC penalty is higher than for the AIC, if we have more than seven observations and the BIC favours simpler models as sample size increases. Although the criteria seem to be very similar, their motivation is grounded in very different principles. While the AIC arises from information theory and tries to find the model with the smallest distance to a complex true model, the BIC is related to an asymptotic Bayes factor and assumes that true model is contained in available set of models (Schwarz, 1978). However, as pointed out by Anraku (1999), none of these criteria is suitable in our framework, since they ignore order restrictions.

The order restricted information criterion (ORIC, Anraku, 1999) uses an order restricted likelihood in which the mean response at each dose level is estimated using isotonic regression (Barlow *et al.*, 1972) and a penalty term is given by

$$\tau(ORIC) = 2 \cdot \sum_{\ell=1}^{K} \ell P(\ell, K, \boldsymbol{v}). \tag{4.5}$$

The level probabilities, $P(\ell, K, \boldsymbol{v})$, represent the probability under the null model (of no dose effect, i.e. under g_0) that number of unique values of dose-specific means μ_i (i.e. number of different dose means) equals to ℓ, while there are K doses for an experiment with a control and $K-1$ dose levels (Robertson *et al.*, 1988). The weights are given by $v_i = n_i/\sigma_i$ and they are constant for balanced experiment with equal variances. The generalized ORIC (GORIC, Kuiper *et al.*, 2011) is an extension for more complicated profiles than simple order restrictions. The GORIC uses maximum likelihood estimate under given constraints and generalizes penalty term. In our framework, for normally distributed data and monotonicity, the GORIC reduces back to the ORIC.

Within the hierarchical Bayesian framework, deviance information criterion (DIC, Spiegelhalter *et al.*, 2002) is often used for model selection. For the DIC, the goodness of fit is measured by $-2\log L(\mathrm{data}|\bar{\boldsymbol{\theta}})$, which is the likelihood of the observed data evaluated using the posterior mean of $\boldsymbol{\theta}$. The penalty for complexity, τ, equals to $\tau = 2p_D$, where p_D is the effective number of parameters of the model. According to Spiegelhalter *et al.* (2002), p_D is a difference between posterior mean of the deviance and deviance evaluated in posterior means of the parameters. Alternatively, Gelman *et al.* (2004) defines p_D as half of variance of the deviance. This estimate shows robustness and accuracy and it is not affected by reparametrization of the model (Spiegelhalter *et al.*, 2014).

The weights defined in Equation (4.3) can be used to estimate the dose-specific means as weighted average of the means estimated by the $R+1$ candidate models. This approach is closely related to model averaging techniques as discussed, in the context of dose-response modelling, in Bretz *et al.* (2005), Pinheiro *et al.* (2006), Whitney and Ryan (2009) and Lin *et al.* (2012c).

Note that it is necessary to fit all candidate models $g_0, \ldots, g_R$ in order to compute the weights based on the IC described in this section. Therefore, with an increasing number of candidate models (e.g. when the number of dose levels increases), the number of fitted models increases as well.

4.2.2 Order restricted estimation: hierarchical Bayesian approach

As discussed in Chapter 2, we can formulate an order restricted Bayesian hierarchical model to estimate the means. As explained in Section 2.3, in order to ensure monotonicity among the means, the prior distributions of all components of vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{K-1})$ are truncated (at zero) normal distributions. Note that $P(\theta_h = 0) = 0$, a probability of any of the components to be exactly zero is equal to zero. Hence, the parametrization in Equation (2.15) implies that a Bayesian inequality model, i.e. a model with $K-1$ (ordered) parameters θ_ℓ, is fitted (Dunson and Neelon, 2003). For example, for $K = 4$, only model g_7 can be fitted. Therefore, necessarily $\mathrm{MED} = 1$. However, all the other models $g_0, \ldots, g_{R-1}$ can be fitted by a slight modification of the parametrization of the mean structure, i.e. by fixing appropriate θ_ℓ to be equal to zero. The DIC, can be used to select the best model and to determine MED, as described in previous section. This approach, however, shares the disadvantage of all IC based methods, its necessity to fit all the models separately.

4.2.3 BVS model approach

The Bayesian variable selection (BVS) model, discussed in details in Section 2.4, is an extension of Bayesian inequality model and it allows us to fit all candidate models at once (through one MCMC chain) via the internal variable selection procedure. As mentioned in Section 2.4 the configuration of the latent variable z determines uniquely all candidate order restricted one-way ANOVA models. Therefore, it gains a clear advantage over any IC based method, where all the models need to be fitted separately.

The posterior mean of z_h, see Equation (2.22), represents the posterior inclusion probability of δ_h in the model (O'Hara and Sillanpää, 2009). Due to the fact that the configuration of the vector $\boldsymbol{z}$ determines unambiguously a particular model, the posterior probability of a particular configuration of $\boldsymbol{z}$ translates into posterior probability of a particular model (Table 2.1). For example, in case of $K = 4$, posterior probability of model g_1 equals to

$$P(g_1|\mathrm{data}) = P\left[\boldsymbol{z} = c(1,0,0)|\mathrm{data}\right]. \tag{4.6}$$

Note that $P(g_r|\mathrm{data})$ is interpreted as posterior probability of model g_r, given the data, the priors and the set of all models. Naturally, prior specification can strongly influence the results of the analysis. In this way, prior information allows us to include information coming from scientific knowledge or previous experiments. Although we usually apply the BVS in case that all models are of interest (e.g. all models from Table 2.1), if a subset of the models is a priori considered impossible, it can be easily omitted by setting its prior probabilities to zero. In case of lack of any prior information, non-informative priors can be used instead, as in Equation (2.21).

Analogously to the previous section, the MED can be obtained by summing the posterior probabilities of appropriate models. The resulting quantities represent the posterior distribution of the MED, i.e. to each possible value of the MED the posterior probability of being the true underlying MED is assigned. For example, for $K = 4$, $P(MED = 2|\text{data}) = P(g_2|\text{data}) + P(g_6|\text{data})$. Hence, in terms of the inclusion vector $\boldsymbol{z}$, the posterior the posterior probability is given by

$$\bar{P}(MED = 2|\text{data}) = \bar{P}(\boldsymbol{z} = (0,1,0)|\text{data}) + \bar{P}(\boldsymbol{z} = (0,1,1)|\text{data}). \tag{4.7}$$

As shown in Section 2.4, the posterior model probabilities play an important role in the estimation of the dose-specific means as well. The means are computed as averaged estimates of means under specific model weighted by posterior probability of that model, that is

$$\bar{\boldsymbol{\mu}} = \sum_{r=1}^{R} \bar{P}(g_r|\mathrm{data})\hat{\boldsymbol{\mu}}_{g_r}. \tag{4.8}$$

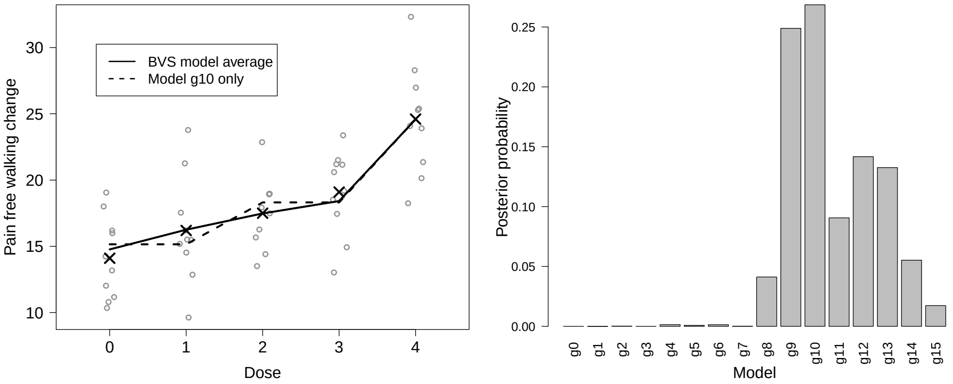


Figure 4.1: *The Angina data. Left panel: Observed data, sample means (crosses) and posterior means of the BVS model (solid line) and model g_{10} (dashed line). Right panel: Posterior probability for g_r, $r = 0, \ldots, 15$. Notation corresponds to the model numbers presented in Table 6.2, extended respectively for $K = 5$ (see Table 6.2).*

4.3 Results

We apply the BVS model, the GORIC, the AIC and the BIC methods for the Toxicity and the Angina data sets described in Section 1.1. The attention is given to the comparison between the BVS and the GORIC, since they are both taking into account order constraints within the estimation procedure of the MED. The model weights based on the IC are interpreted (in terms of Equation 4.3) as posterior model probabilities. In order to distinguish between the results of the methods, we denote posterior probabilities as $\bar{P}_{GORIC}$ and $\bar{P}_{BVS}$ for respective method. The analysis for all methods was done using the R software (R Core Team, 2014) version 3.1.1. For the BVS model, the MCMC was run using the package runjags (Denwood, In Review) together with the JAGS software (Plummer, 2003).

The results for the BVS model are shown in Figure 4.1 and Figure 4.2 for the Angina data and the Toxicity data, respectively. The left panels show the data, the BVS weighted average of mean estimates (solid line) and the best model selected by BVS (dashed line). For both case studies, the effect of model averaging is clearly seen. The right panels of both figures show the posterior model probabilities. While there is much clearer candidate for the best model for Toxicity data, g_1 with $\bar{P}_{BVS}(g_1|data) = 0.38$, the result for Angina data supports nearly equally two models, g_9 ($\bar{P}_{BVS}(g_9|\text{data}) = 0.249$) and g_{10} ($\bar{P}_{BVS}(g_{10}|\text{data}) = 0.269$).

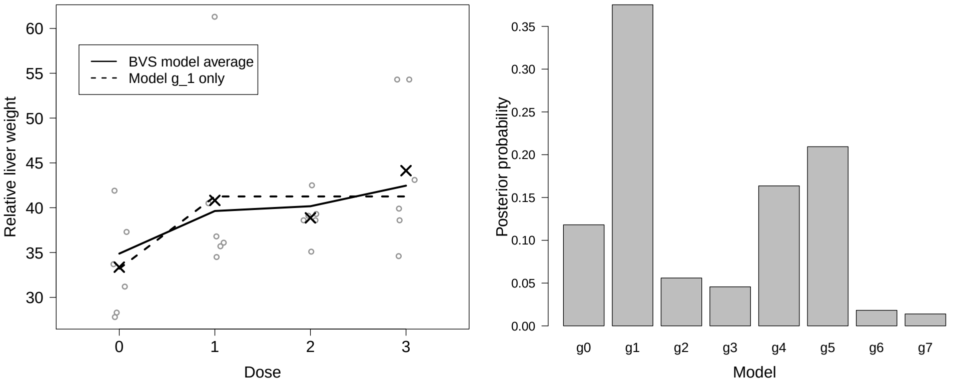


Figure 4.2: *The Toxicity data. Left panel: Observed data, sample means (crosses) and posterior means of the BVS model (solid line) and model g_1 (dashed line). Right panel: Posterior probability for g_r, $r = 0, \ldots, 7$. Notation corresponds to the model numbers presented in Table 2.1.*

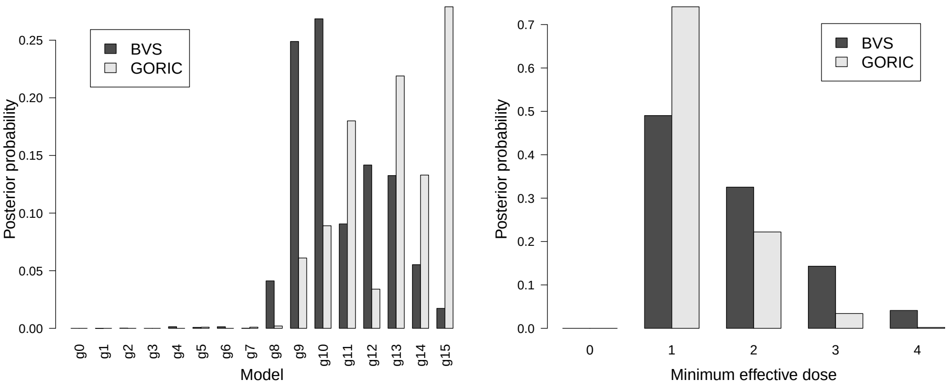


Figure 4.3: *The Angina data. The BVS results (black) and GORIC results (grey) comparison. Left panel: Posterior probability for g_r, $r = 0, \ldots, 15$. Right panel: Posterior probability for the MED.*

The posterior model probabilities obtained for the BVS and the GORIC for the Angina data set are shown in left panel of Figure 4.3. For both methods, the highest posterior probabilities were obtained for models with an increment between the last two doses. However, the GORIC tends to prefer more complex models with smaller increments across multiple doses (g_{13}, g_{15}), while the BVS selects models with just few larger increments (g_9, g_{10}). The posterior probabilities of the MED are shown in the right panel of Figure 4.3. Both the GORIC and the BVS assigned the highest posterior probability of being MED to the first dose. However, there is a difference between the two methods. Since the GORIC method selects models with more parameters, it gives higher probability to models with increment already between first and second dose and therefore $P_{GORIC}(\text{MED} = 1|\text{data})$ is estimated with large posterior probability, $\bar{P}_{GORIC}(\text{MED} = 1|\text{data}) = 0.741$. It also assigns nearly zero probability to $\bar{P}_{GORIC}(\text{MED} = 4|\text{data}) = 0.002$. In contrast, the BVS method gives much lower posterior probability to $\bar{P}_{BVS}(\text{MED} = 1|\text{data}) = 0.490$ and the posterior distribution of the MED is more equally spread over all doses, i.e. $\bar{P}_{BVS}(\text{MED} = 2|\text{data}) = 0.325$ and $\bar{P}_{BVS}(\text{MED} = 4|\text{data}) = 0.041$. The complete results are presented in Table 4.1. We can see that the results obtained for the AIC and BIC methods lie between the results obtained for the GORIC and the BVS methods. Note that the results for the BIC are much closer to results of the BVS.

Table 4.1: Estimated posterior model probabilities for the Angina data for GORIC, AIC, BIC and BVS. First column: Order restricted log-likelihood.

Profile	ORLL	GORIC	AIC	BIC	BVS
g_0	-149.77	0.00	0.00	0.00	0.00
g_1	-144.55	0.00	0.00	0.00	0.00
g_2	-141.46	0.00	0.00	0.00	0.00
g_3	-140.80	0.00	0.00	0.00	0.00
g_4	-138.65	0.00	0.00	0.00	0.00
g_5	-136.92	0.00	0.00	0.00	0.00
g_6	-137.39	0.00	0.00	0.00	0.00
g_7	-136.61	0.00	0.00	0.00	0.00
g_8	-135.97	0.00	0.01	0.04	0.04
g_9	-132.31	0.06	0.13	0.21	0.25
g_{10}	-131.99	0.09	0.18	0.29	0.27
g_{11}	-131.01	0.18	0.17	0.11	0.09
g_{12}	-133.01	0.03	0.06	0.11	0.14
g_{13}	-130.82	0.22	0.21	0.13	0.13
g_{14}	-131.42	0.13	0.12	0.07	0.06
g_{15}	-130.43	0.28	0.11	0.03	0.02

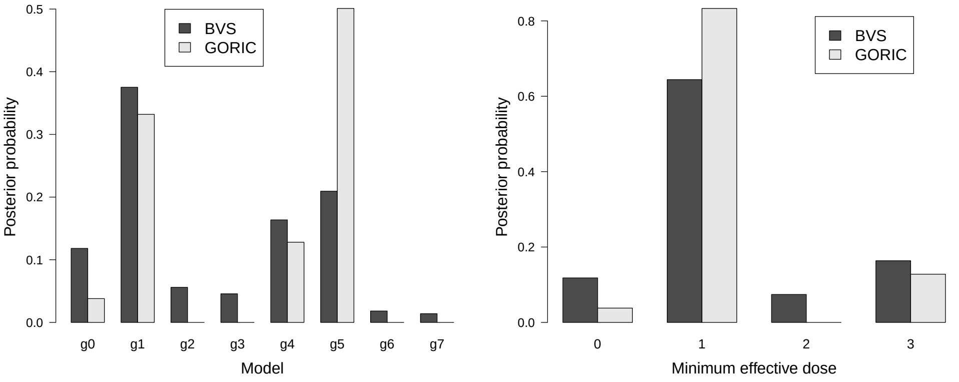


Figure 4.4: *The Toxicity data. The BVS results (black) and GORIC results (grey) comparison. Left panel: Posterior probability for g_r, $r = 0, \ldots, 7$. Right panel: Posterior probability of the MED.*

Similar pattern can be seen for the Toxicity data in Figure 4.4. While the GORIC prefers a more complex model g_5 (having three different means) with $\mathrm{MED} = 1$, the BVS suggests that the best model is g_1, while giving much higher posterior probabilities to other models, such as g_0, g_4 and g_5. Once again, both methods estimated the highest posterior probability of being the MED for the same dose level, with the GORIC estimate $\bar{P}_{GORIC}(\mathrm{MED} = 1|\mathrm{data}) = 0.833$ and the BVS estimate $\bar{P}_{BVS}(\mathrm{MED} = 1|\mathrm{data}) = 0.644$. Similarly to the Angina data, the GORIC assigns very high posterior probability to $\mathrm{MED} = 1$ (see right panel of Figure 4.4), while BVS spread probability more equally, estimating relatively high posterior probabilities for other doses. Note that in Table 4.2 not all models were fitted for the GORIC, AIC and BIC. That is caused by the violation of monotonicity assumption in the observed means between dose 2 and dose 3 (see Figure 4.2). As mentioned above, isotonic regression was used to estimate the order restricted means. While we incorporate the order restrictions for maximum likelihood estimation, the models with increase between dose 2 and dose 3 reduced to models that have a flat mean profile between dose 2 and dose 3 (e.g. model g_2 will reduce to model g_0). Therefore, only a subset of models g_0, g_1, g_4, g_5 with no increment between the dose 2 and dose 3 can be actually fitted and estimated. This property does not apply to the BVS model, because it does not use isotonic regression for the estimation of the means.

In both data sets, the GORIC seems to support models with less equalities (i.e. more complex models) compared to the BVS and therefore estimates the lower values of the MED with higher probabilities. Both methods tend to select similar patterns, but small

Table 4.2: Estimated posterior model probabilities for the Toxicity data for GORIC, AIC, BIC and BVS. First column: Order restricted log-likelihood. Note that, as explained in Section 4.3, some of the models were not fitted for IC; due to the incorporated order restrictions they reduced to other models.

Profile	ORLL	GORIC	AIC	BIC	BVS
g_0	-82.98	0.04	0.08	0.16	0.12
g_1	-80.32	0.33	0.42	0.46	0.38
g_2	—	0	0	0	0.06
g_3	—	0	0	0	0.05
g_4	-81.28	0.13	0.16	0.18	0.16
g_5	-79.51	0.50	0.34	0.21	0.21
g_6	—	0	0	0	0.02
g_7	—	0	0	0	0.01

differences between consecutive doses are treated as flat by the BVS but as increments by the GORIC. The cause of this difference is due to the fact that the penalty of GORIC is rather low when additional parameters are added to the model. Hence, the GORIC supports more complex models and results in much higher $\bar{P}_{GORIC}(\mathrm{MED} = 1|\mathrm{data})$. On the other hand, the results for the BVS suggest that a model reduction step is addressed automatically within the procedure and a relatively large difference among doses is needed to include the increment in the model. As a consequence, the distribution of $\bar{P}_{BVS}(\mathrm{MED} = i|\mathrm{data})$ is spread more equally across the doses. The AIC and BIC are somewhere between the other two methods, AIC being closer to GORIC and BIC closer to BVS. This is expected since compared to the AIC, the BIC has a tendency to select less complex models due to a high penalty term. The values of penalties for Angina data set are shown in Table 4.3 and for Toxicity data set in Table 4.4 (note that we list only the models that were possible to fit for this particular data set).

As expected, the choice of the criterion determines the posterior distribution of MED. Although the MED with the highest posterior probability could be the same for different methods, substantial differences can be observed in the underlying posterior distribution that quantifies the uncertainty in the choice of MED. On the other hand, the choice of the criterion can incorporate our preference for a more or less complex model in the process of the estimation of the posterior probabilities.

Table 4.3: Penalties for different models fitted for the Angina data. First column: Order restricted log likelihood. Remaining columns: Penalty term for respective IC.

Profile	ORLL	GORIC	AIC	BIC
g_0	-149.77	2.00	4	7.82
g_1	-144.55	2.50	6	11.74
g_2	-141.46	2.50	6	11.74
g_3	-140.80	2.79	8	15.65
g_4	-138.65	2.50	6	11.74
g_5	-136.92	2.86	8	15.65
g_6	-137.39	2.77	8	15.65
g_7	-136.61	3.03	10	19.56
g_8	-135.97	2.50	6	11.74
g_9	-132.31	2.92	8	15.65
g_{10}	-131.99	2.86	8	15.65
g_{11}	-131.01	3.14	10	19.56
g_{12}	-133.01	2.79	8	15.65
g_{13}	-130.82	3.14	10	19.56
g_{14}	-131.42	3.03	10	19.56
g_{15}	-130.43	3.28	12	23.47

Table 4.4: Penalties for different models fitted for the Toxicity data. First column: Order restricted log likelihood. Remaining columns: Penalty term for respective IC.

Profile	ORLL	GORIC	AIC	BIC
g_0	-82.98	2.00	4	6.36
g_1	-80.32	2.50	6	9.53
g_4	-81.28	2.50	6	9.53
g_5	-79.51	2.89	8	12.71

4.4 Simulation study

4.4.1 Simulation setting

Considering the findings in Section 4.3, we conducted a simulation study to explore suitability of various methods according to true underlying model. The data were generated according to the model order restricted one-way ANOVA model specified in Equation (2.4), $Y_{ij} \sim N(\lambda\mu_i, \tau^{-1})$, with $\tau = 1$ and varying λ. The parameter λ is used to control the magnitude of increment of the mean response from one dose level to another. The higher the value of λ, the higher the increment. The simulation represented an experiment with $K = 4$ dose levels and $n = 3$ observations per dose and followed the design described in Section 6.1. Magnitude of the dose-response effect was represented by varying parameter $\lambda = 1, 2, 3$ (see Section 6.1). In total, $N = 1000$ data sets were generated for each combination of a specific model and λ (i.e. in total 22 combinations were simulated, 7×3 for $g_1, \ldots, g_7$ and one for g_0, each 1000 times).

For all the methods, an assumption of a non-decreasing trend was made. As explained in the previous section, not all the models can be fitted for the IC methods in each simulated data set (when violation of monotonicity in simulated means occurs), while the BVS provided posterior probability for all the models in each simulated data set. The posterior model probabilities, $\bar{P}(g_r|\mathrm{data})$, were computed according to the BVS, AIC, BIC and GORIC methods. The posterior probabilities for the MED, $\bar{P}(\mathrm{MED} = i|\mathrm{data})$, were derived by summation of appropriate posterior model probabilities. The methods were evaluated based on two criteria: the correct identification of the true underlying model and the correct identification of the true underlying MED. Additionally, the setting when the best model and the second best model are considered for evaluation is briefly discussed in Section 4.5 and the full results are shown in Section 6.4.

4.4.2 Simulation results

As shown in Table 4.5, performance according to model complexity is profound in simulation study results. While the BVS clearly performs better for simple models with only one or two different mean levels (g_0, g_1, g_2 and g_4), the GORIC achieves better results for complex models (g_3, g_6, g_7). The result for model g_5 highlights another interesting point. While the magnitude of the difference is getting higher, the GORIC seems to prefer more complex models (splitting high increment among more dose levels). Therefore, if $\lambda = 3$, the BVS overtakes the GORIC in terms of correct selection of the model g_5 and reduces the difference for models g_3 and g_6. Clearly, the GORIC is better method for the detection of model g_7. On the other hand, it shows the worst performance for the simplest model g_0

that can be of profound interest, representing absence of dose-response relationship. Interestingly, the AIC method performs well. While being always between BVS and GORIC, it shows good performance, except for model g_7. Performance of BIC is rather poor, being among the worst methods for all the possible models (and except g_0, being always worse than AIC). The complexity of the models selected by a specific method depends on the penalty term of that method. Typically, it holds that penalty of the GORIC is smaller than penalty of the AIC that is (for $n > 7$) smaller than penalty of the BIC. Therefore, the AIC and GORIC may select more complex models. The AIC and GORIC methods arise from information theory and they estimate Kullback-Leibler divergence (Kullback and Leibler, 1951) between the true model and models under consideration. Therefore, they do not assume that the true model is necessarily among the candidate models and they try to approximate it. In contrast, the BVS model selects the best model among the candidate models. Additional results for varying number of replicates within dose ($n = 4, 5, 10$) indicate the same patterns and are presented in Section 6.4.

The main goal of the analysis is to estimate the MED. The evaluation of methods based on correct identification of the MED, presented in Table 4.6, leads to different conclusions than correct model selection based analysis. We can see an overall improvement in the correct identification rate. This is due to the fact that if the true model is not selected, the methods tend to select the model with the same MED. The clearest improvement occurs for the GORIC, especially for model g_1. The magnitude of the increment, represented by λ, seems to be an important factor for a correct MED determination. Clearly, the GORIC performs better for $\lambda = 1$ for most of the models, while the BVS outperforms the GORIC for nearly all of the models if $\lambda = 3$. The model complexity factor stays clearly visible only for model g_4 (increment only in last dose) and g_7 (increment in all doses). The AIC seems very suitable for MED selection. It has never been the best method, but it has never had worse performance than both BVS and GORIC simultaneously. The BIC does not provide good results, in some cases it performed slightly better than other methods, but it is often the worst method with rather poor overall performance. Similar results for additional settings are presented in Section 6.4.

Table 4.5: Comparison of estimated probability of the correct model selection based on 1,000 simulated data sets for BVS, GORIC, AIC and BIC criterion for $K = 4$, $n = 3$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.73	0.59	0.76	0.81
1	g_1	0.57	0.51	0.53	0.49
	g_2	0.46	0.42	0.47	0.46
	g_3	0.03	0.16	0.05	0.03
	g_4	0.55	0.48	0.51	0.48
	g_5	0.08	0.22	0.09	0.07
	g_6	0.02	0.16	0.04	0.02
	g_7	0.00	0.03	0.00	0.00
2	g_1	0.83	0.63	0.78	0.80
	g_2	0.78	0.54	0.73	0.77
	g_3	0.22	0.48	0.30	0.23
	g_4	0.82	0.61	0.78	0.79
	g_5	0.43	0.54	0.49	0.42
	g_6	0.23	0.46	0.29	0.24
	g_7	0.01	0.28	0.04	0.02
3	g_1	0.88	0.63	0.79	0.83
	g_2	0.84	0.55	0.76	0.81
	g_3	0.59	0.66	0.64	0.60
	g_4	0.86	0.62	0.80	0.83
	g_5	0.79	0.67	0.77	0.77
	g_6	0.57	0.65	0.63	0.59
	g_7	0.09	0.62	0.25	0.19

Table 4.6: Comparison of estimated probability of a selection of the correct MED based on 1,000 simulated data sets for BVS, GORIC, AIC and BIC criterion with $K = 4$, $n = 3$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.73	0.59	0.76	0.81
1	g_1	0.62	0.73	0.61	0.55
	g_2	0.47	0.51	0.49	0.47
	g_3	0.40	0.53	0.39	0.34
	g_4	0.55	0.48	0.51	0.48
	g_5	0.39	0.53	0.39	0.35
	g_6	0.32	0.40	0.36	0.34
	g_7	0.32	0.44	0.32	0.29
2	g_1	0.96	0.99	0.96	0.94
	g_2	0.83	0.72	0.82	0.83
	g_3	0.61	0.81	0.65	0.59
	g_4	0.82	0.61	0.78	0.79
	g_5	0.70	0.85	0.74	0.71
	g_6	0.57	0.60	0.59	0.59
	g_7	0.48	0.70	0.53	0.48
3	g_1	1.00	1.00	1.00	1.00
	g_2	0.91	0.72	0.86	0.90
	g_3	0.82	0.94	0.86	0.83
	g_4	0.86	0.62	0.80	0.83
	g_5	0.90	0.98	0.93	0.91
	g_6	0.75	0.69	0.76	0.76
	g_7	0.64	0.86	0.71	0.66

4.5 Discussion

This chapter discusses the Bayesian variable selection method for the model selection and the estimation of the minimum effective dose. A comparison with competing methods based on information criteria GORIC, AIC and BIC was conducted in both case studies and simulation study. One advantage of BVS is its unified framework for inference, estimation and model selection. While posterior probabilities $\bar{P}(g_r|\mathrm{data}, g_0, \ldots, g_R)$ can be used as a model selection tool, the dose-specific means estimates are based on the weighted average of model-specific estimates according to the posterior model probabilities. As shown in Chapter 3, the posterior probability of the null model can be used for inference. Therefore, the BVS model provides estimates for the dose-specific means while taking model uncertainty into account. Similarly, the model averaged estimates can be obtained for IC methods by using model-specific maximum likelihood estimates weighted by appropriate model weights.

In terms of model and MED selection, the main advantage of BVS is that it fits all the models simultaneously. It is not necessary to check which model can be actually fitted to the data due to violations of monotonicity. In contrast with the IC based methods, the number of fitted models does not increase with increasing number of dose levels. For $K > 5$, the amount of models to be fitted can become prohibitive for IC based methods that require to fit all candidate models separately in order to estimate the posterior model probabilities.

The isotonic regression procedure used for all IC based methods raises an important issue. As we have seen for the Toxicity data set, not all the models can be fitted due to violation of monotonicity assumption. Analogously, we have seen in the simulation study that it may happen (for single experiment) that the true model that generated the data may not be fitted due to the variability in the data. Hence, the posterior probability of true model may be zero for all IC based methods. Naturally, this issue carries over to the MED estimation as well. The true MED can be missed, if the variability in the data causes the violation of monotonicity in the dose-specific means.

The BVS model outperforms the other methods in case of less complex underlying models or higher magnitude of overall difference. In case of small differences, it tends to oversimplify the models, especially for the most complex model g_7. On the other hand, the GORIC method prefers complex models, leading to its poor performance in case of high magnitude of difference and simplest models as g_0 or g_1. While taking into account not only the best model, but also the second best model (with respect to posterior probability), the BVS model performs much better, relatively to IC methods (additional simulations supporting this claim are presented in Section 6.4). In particular, model g_0

can be interpreted in terms of the null hypothesis of no dose-response relationship. The low sensitivity of the GORIC for this model suggests that the method should be used only after an initial inference step. The performance for higher magnitudes is of main interest from an application point of view. As mentioned in the Introduction, the MED is typically related to the clinical significance as well as to the statistical significance. Therefore, cases of small overall effects are not of imminent interest. The bigger the overall dose effect is, the higher is the chance that MED would be relevant and its correct estimate is needed.

Chapter 5

Robustness Against the Prior Configuration and Model Complexity

5.1 Introduction

The aim of this chapter is to investigate influence of the choice of the prior model probabilities on the estimation of the dose-specific means, model selection procedures and the inference. Additionally, we define the model complexity within the BVS framework as posterior expected complexity and we present an investigation about its properties. This measure is analogous to the penalty for complexity used by information criteria (e.g. AIC, BIC, DIC, etc.). Within the Bayesian framework, it represents the expected number of parameters of the true model, given the model uncertainty. Each model has a known number of distinct dose-specific means and posterior model probability. Hence, a weighted average of these quantities results in posterior complexity of the set of the models. Analogously, the "prior complexity" can be obtained using prior model probabilities. There is a clear link between the BVS model complexity and the likelihood-ratio test (LRT) and the order restricted information criterion. All the quantities are related to the level probabilities, i.e. probabilities of having certain number of levels under the null model. This topics will be explained further in Section 5.2.

Clearly, estimation, model selection, inference and model complexity depend, to some extent, on the configuration of the prior model probabilities. As explained in Chapter 3, we use for inference a permutation based procedure. In the first part of this chapter, we investigate how sensitive is the inference to the choice of the prior model probabilities. Note that we expect that the results of the inference procedure will not change when the set of prior model probabilities changes. The reason is that permutation is done conditionally on the prior, so the effect of the priors should be diminished when enough permutations are conducted. As a consequence, the inference is expected to be robust against the configuration of non-informative priors. This is desired property in case that there is no prior information and non-informative priors are chosen, as will be explained further. In contrary, model selection or the estimation of the minimum effective dose (MED) should be strongly influenced by the choice of prior distribution, because both procedures directly rely on posterior model probabilities. Dependency of the posterior probabilities on the prior specification is related to amount of information in the data. Especially in case of small sample size or relatively large noise, the posterior probabilities will be dominated by the prior probabilities of the models. The behaviour of estimates for the dose-specific posterior means is not clear a priori and is a subject for investigation in this chapter. Although the estimates depend on the posterior probabilities of the models, the model averaging process, discussed in Chapter 2, could compensate for the effect of prior model probabilities. Even if the "true" model is not correctly identified, models that are rather close to it could be selected instead. Therefore, the weighted average of the models could still provide an accurate estimate of underlying dose-specific means. Similarly, the complexity measure is based on a weighted average of the posterior model probabilities, so it is not a priori clear, how much it would be influenced by configuration of priors.

The methodological background of this chapter is described in detail in Section 5.2. A motivating example in which the proposed BVS model with varying priors is applied to the case study is shown in Section 5.3 and the influence on estimation, model selection, inference and posterior expected complexity is evaluated. In Section 5.4, a simulation study is conducted to investigate the influence of the choice of priors on the different aspects of the BVS framework. Finally, a discussion is given in Section 5.5.

5.2 Methodology

Analogously to the previous chapters, let $\boldsymbol{Y} = \{Y_{ij},\ \ i = 0, \ldots, K-1,\ \ j = 1, \ldots, n_i\}$ denote the set of the observations, where n_i represents the sample size at dose i (i.e. j represents the replicates within the dose). Further, it is assumed that the dose-specific means $\mu_0, \ldots, \mu_{K-1}$ follow a simple order, i.e. a monotone order of the form $\mu_0 \leq \cdots \leq \mu_{K-1}$ or $\mu_0 \geq \cdots \geq \mu_{K-1}$, for an upward and downward trends, respectively. The aim is to model the relationship between dose and the response of interest, while accounting for model uncertainty. Therefore, different types of profiles of dose-response relationship are determined by the presence of equality or inequality between consecutive dose-specific means, resulting in the set of $\{g_0, \dots, g_R, R = 2^{K-1}\}$ one-way order restricted ANOVA models for control and $K-1$ active dose levels. The model g_0 represents null hypothesis of equality of dose-specific means (see Equation 2.1 in Section 2.1). The union of all the remaining models represents alternative hypothesis of at least one strictly monotone relationship (see Equation 2.2). For example, for the dose-response experiment with control dose and three increasing dose levels (i.e. $K = 4$), there are eight possible one-way order restricted ANOVA models presented in Table 2.1.

5.2.1 Level probabilities

The estimation of dose-specific means under monotonicity assumption using the maximum likelihood estimators leads to the isotonic regression (Barlow *et al.*, 1972). Denote isotonic means as $\hat{\mu}_0^* \leq \dots \leq \hat{\mu}_{K-1}^*$. They can be estimated from dose-specific sample means $\hat{\mu}_0, \dots, \hat{\mu}_{K-1}$ using the 'pool adjacent violators algorithm' (PAVA). In first step, initiate $\hat{\mu}_i^* = \hat{\mu}_i$ for all $i = 0, \dots, K-1$. Afterwards, for any pair j and $j+1$ for which the order is violated, i.e. $\hat{\mu}_j^* > \hat{\mu}_{j+1}^*$, the isotonic means are updated as $\hat{\mu}_j^* = \hat{\mu}_{j+1}^* = (n_j\hat{\mu}_j^* + n_{j+1}\hat{\mu}_{j+1}^*)/(n_j + n_{j+1})$. The procedure is repeated until all the means comply with monotone order restriction.

The level probabilities represent the probabilities of obtaining certain number of unique isotonic means, i.e. 'levels', if isotonic regression is applied to the data generated under the null hypothesis. Let us denote $P(\ell, K, \boldsymbol{w})$ level probability of obtaining ℓ levels for K dose levels, while the inverse of $\boldsymbol{w}$, $\boldsymbol{w}^{-1} = (w_0^{-1}, ..., w_K^{-1})$, consist of variances of the response at each dose. For example, for case of $K = 4$ and equal weights $\boldsymbol{w}_0$, the probability for one single level are equal to $P(\ell = 1, 4, \boldsymbol{w}_0) = 0.25$, $P(\ell = 2, 4, \boldsymbol{w}_0) = 0.46$, $P(\ell = 3, 4, \boldsymbol{w}_0) = 0.25$ and $P(\ell = 4, 4, \boldsymbol{w}_0) = 0.04$ (Robertson *et al.*, 1988 and Shkedy *et al.*, 2012a). The last probability implies that the data generated under the null rarely induce strictly monotone sequence of dose-specific means. More likely, two unique isotonic means may occur which corresponds to the oscillation of dose-specific means around true underlying mean. The level probabilities refer only to the number of unique estimates for the means and not to the significance of the difference between particular means. Therefore, they are independent on the variability of the data as far as the variance is constant across all the doses.

The importance of level probabilities is obvious in order restricted setting, both in inference and model selection framework. Consider the LRT of null hypothesis against an ordered alternative. The p-value for the test is given by (Barlow *et al.*, 1972)

$$P_{H_0}(T_{LRT} \geq t_{LRT}) = \sum_{\ell=1}^{K} P(\ell, K, \boldsymbol{w}) P\left[B_{\frac{1}{2}(\ell-1),\frac{1}{2}(N-\ell)} \geq t_{LRT}\right]. \tag{5.1}$$

Here, N is the total number of observations. $B_{\frac{1}{2}(\ell-1),\frac{1}{2}(N-\ell)}$ denotes Beta distribution with $\alpha = 1/2(\ell - 1)$ and $\beta = 1/2(N - \ell)$ and $B_{0,\beta} \equiv 0$. The higher number of levels under the null implies generally higher values of T_{LRT} statistic. Therefore, the overall distribution of the test statistics is a mixture of Beta distributions weighted by level probabilities (Shkedy *et al.*, 2012a).

Similarly, level probabilities play an important role in likelihood based approaches for the model selection via information criteria. As shown in Chapter 4, an order restricted information criterion, ORIC (Anraku, 1999) can be used for model selection. The ORIC is derived from Kullback-Leibler divergence (Kullback and Leibler, 1951) minimization accounting for order restriction. The criterion uses the order restricted likelihood that is related to isotonic regression to measure the goodness of fit of the model. The level probabilities are used in the penalty term as

$$ORIC = -2\text{log}L(\boldsymbol{\theta}|\text{data}) + 2 \cdot \sum_{\ell=1}^{K} \ell P(\ell, K, \boldsymbol{w}). \tag{5.2}$$

The weights $w_i = n_i/\sigma_i$ are constant for balanced experiment with equal variances. The use of level probabilities naturally reflects the model fitting via isotonic regression. The penalty of IC depends on a number of distinct parameters. In case of the order restricted framework, the number of parameters under H_0, i.e. number of unique isotonic means, varies based on the number of violations of monotonicity among the sample means. The level probabilities express the probabilities of obtaining number ℓ of distinct means, so their weighted average translates to expected number of distinct isotonic means for given model (under H_0). There is a clear analogy between the ORIC and the AIC that simply takes as penalty term the number of parameters in the model.

5.2.2 Posterior expected complexity

In previous section, the penalty term of the ORIC is expressed in terms of an expected number of levels, i.e. distinct isotonic means in the isotonic regression solution. Following Equation (5.2), it is clear that the expected complexity (EC) equals to weighted sum of number of levels:

$$EC = \sum_{\ell=1}^{K} P(\ell, K, \boldsymbol{w}) \cdot \ell. \tag{5.3}$$

For example, in case of $K = 4$, the result of $EC = 2.083$ suggests that we expect two distinct isotonic means prior to looking at the data in case that the data were generated under the null distribution. The level probabilities are computed under the null hypothesis, when there is only one level of mean. However, variability in the sample means caused by the noise in the data results in varying number of observed levels. For example, for a small simulation in which 1,000 data sets with $K = 4$ and $n = 3$ were generated under the null hypothesis and the means were estimated using isotonic regression, there were 226 experiments with one level only, 486 with two unique levels, 259 with three levels and 29 experiments with four unique levels. As expected, the mean number of levels across all data sets was 2.091, close to $EC = 2.083$. Note that the estimated rate of the numbers of levels (e.g. 0.226 for one level) is close to theoretical level probabilities for $K = 4$ mentioned in Section 5.2.1 (see Figure 5.1). As mentioned above, the EC is the expected number of levels when isotonic regression is used to estimate the means and the data are generated under the null hypothesis, i.e. the number of levels of the underlying true model is one $(pNL = 1)$, because the true model has exactly one dose-specific mean.

Within the Bayesian framework and under the model uncertainty, the generalized prior expected complexity, pEC_0, can be defined as weighted average of the prior probabilities of the models $g_0, \ldots, g_R$ and their corresponding number of levels $\ell_0, \ldots, \ell_R$. The word 'expected' arises from the fact that pEC_0 does not represent the number of levels that can be actually observed in an experiment, but the average number of levels given the prior probabilities. When the prior knowledge is combined with the data, the posterior expected complexity pEC, can be obtained analogously. It can be defined as the sum over the possible models $g_0, \ldots, g_R$ that weights the number of levels ℓ_{g_r} of these models with posterior model probabilities:

$$pEC = \sum_{r=0}^{R} P(g_r|\text{data}) \cdot \ell_{g_r} = \sum_{\ell=1}^{K} P(\ell|\text{data}) \cdot \ell. \tag{5.4}$$

The second sum corresponds to the definition in Equation (5.3), where $P(\ell|\text{data})$ is the sum of posterior probabilities $P(g_r|\text{data})$ of the models that have exactly ℓ levels. The pEC reflect both the data and prior knowledge about the model probabilities. Therefore, it represents posterior complexity, while pEC_0 represents prior complexity.

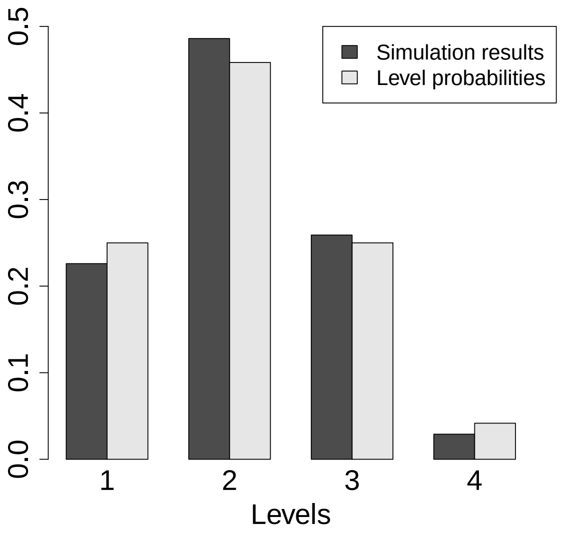


Figure 5.1: *Distribution of number of levels in 1,000 simulated data sets (dark) compared to theoretical level probabilities (light grey) for appropriate setting.*

5.2.3 Choice of priors

The priors for the BVS defined in Equation (2.20) implies the same prior for each of the models $g_0, \ldots, g_R$, , i.e. $1/(R+1)$. This set of prior model probabilities is often considered as non-informative priors. Changing the prior distribution of the models can be easily done in practice by changing the prior distribution of $\boldsymbol{z}$ in the BVS model specified in Equation (2.19) (see Section 2.4).

A formal definition for non-informative priors does not exists due to the fact that non-informative priors can be viewed from different perspectives. The first view follows the reasoning of Jeffreys (1961) by giving the same prior probability to any model under consideration (denoted as E_M for 'equal models'). In this case, we assign the prior $1/(R+1)$ to any of the $g_0, \ldots, g_R$ models. This point of view centers on models as most important entities and treats the null model g_0 in same way as the other models. However, such an approach assigns a prior probability of $R/(R+1)$ to the alternative hypothesis, which, if inference is of primary interest, is an informative prior that favors the alternative hypothesis. Therefore, a second configuration can be considered that assigns a prior of $1/2$ to g_0 and distributes the remaining probability over alternative models $g_1, \ldots, g_R$ as $1/2R$ (denoted as E_H for 'equal hypotheses'). As third option, the number of unique levels can be of primary interest, e.g. if the estimation of the minimum effective dose (MED) is the goal. In this case, a prior of $1/K$ distributed over all the models having k unique means, $k = 1, \ldots, K$, creates a non-informative prior with respect to the number of levels in the model (denoted as E_L for 'equal levels'). The last option to be considered is the specification using level probabilities that represent priors under the null model (denoted as L_P for 'level probabilities'). The example of different prior values for $K = 4$ is given in Table 5.1 and visualized in Figure 5.2.

The prior expected complexities pEC_0 for the four choices of prior distribution are equal to $pEC_0(E_M) = 2.5$, $pEC_0(E_H) = 1.857$, $pEC_0(E_L) = 2.5$ and $pEC_0(L_P) = 2.083$. The equality $pEC_0(L_P) = EC$ holds, because level probabilities are used as priors. Note that $pEC_0(E_M) = pEC_0(E_L)$, but $pEC_0(E_L)$ assigns higher prior weights on 'extreme' models, either with very low or very high number of levels. The smallest pEC_0 is observed for E_H, the prior distribution with the highest weight assigned to the null model g_0, the model with the lowest number of levels.

Table 5.1: Different priors configurations for $K = 4$. The model g_0 represents the null model of no dose effect, i.e. having same mean for all doses. Models g_1, g_2, g_4 have two unique means, models g_3, g_5, g_6 three unique means and model g_7 has four unique means.

Model	Eq. models	Eq. hypothesis	Eq. levels	Level prob.	N. of levels
	E_M	E_H	E_L	L_P	
g_0	0.125	0.5	0.25	0.25	1
g_1	0.125	0.071	0.083	0.153	2
g_2	0.125	0.071	0.083	0.153	2
g_3	0.125	0.071	0.083	0.083	3
g_4	0.125	0.071	0.083	0.153	2
g_5	0.125	0.071	0.083	0.083	3
g_6	0.125	0.071	0.083	0.083	3
g_7	0.125	0.071	0.25	0.042	4

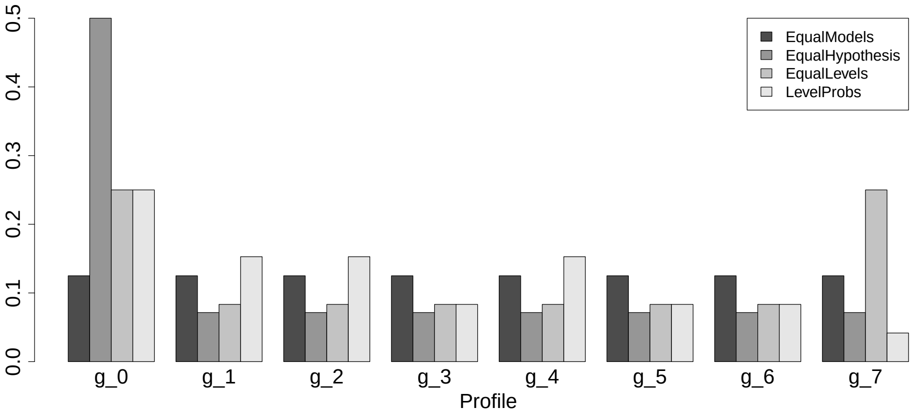


Figure 5.2: *Different priors configurations for $K = 4$. The model g_0 represents the null model of no dose effect, i.e. having same mean for all doses. Models g_1, g_2, g_4 have two unique means, models g_3, g_5, g_6 three unique means and model g_7 has four unique means.*

5.3 Motivating example

We apply the BVS with varying priors to the Toxicity data presented in Section 1.1. The resulting posterior probabilities are shown in Figure 5.3. The choice of priors has clear impact on the posterior probabilities. As expected, $\bar{P}(g_0|\text{data})$ is highest for E_H which assigns the highest prior probability to the null model. Otherwise, model g_1 is preferred, although for E_L, its posterior probability is almost the same as probability of model g_0.

Figure 5.4 shows that the effect on the estimates of the dose-specific means is visible in the first and the last dose. The difference is mostly profound between the configurations in E_M and E_H and it corresponds to the difference between the resulting posterior probabilities of the models g_1, g_4 and g_5, i.e. models that include an increment between the first two and the last two doses. The more robust estimation (with respect to the prior configuration) is observed for the middle doses. The order restriction allows to borrow the information from neighbouring doses and therefore the uncertainty at the bordering doses is much higher than the one in the middle. Analogous behaviour can be observed for any type of regression of continuous variable.

As expected, the permutation test seems to be robust against changes of the prior. The p-values obtained for the different configurations are $p_{E_M} = 0.021$, $p_{E_H} = 0.025$, $p_{E_L} = 0.032$ and $p_{L_P} = 0.032$, respectively. Under all the priors, H_0 is rejected in favour of increasing trend.

Few questions arise now. The results presented in this section suggest that the inference procedure is robust against the configuration of the prior probabilities. Is this the case for this specific data or can it be observed in general? Does the configuration have an effect on the control of Type I error and the power of the test? We have seen that different prior configurations lead to a different posterior expected complexity. How the estimate of the posterior expected complexity changes under different prior configurations and how much it differs from the number of levels obtained from the isotonic regression? For the reminder of this chapter we present a simulation study in which all questions mentioned above are investigated.

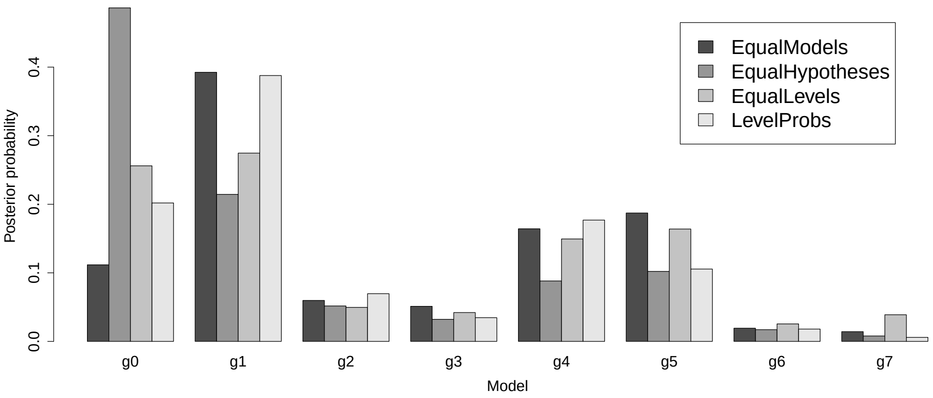


Figure 5.3: *Posterior probabilities for different prior configurations. Each set of bars shows the posterior probability of the particular model $g_0, \ldots, g_7$. Grey scale distinguishes among different prior configurations: darkest to lightest for E_M, E_H, E_L and L_P, respectively.*

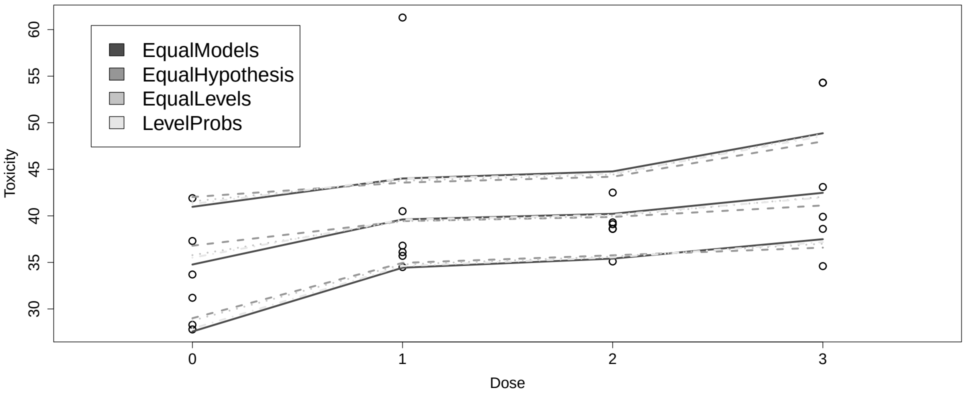


Figure 5.4: *Estimated posterior means with 95% credible intervals for different prior configurations. Grey scale and line types distinguish among different prior configurations: darkest to lightest for E_M (solid), E_H (dashed), E_L (dotted) and L_P (dash dotted), respectively.*

5.4 Simulation study

A simulation study was conducted in order to investigate the influence of the prior configuration on the performance of the permutation based inference procedure for the BVS method in terms of the Type I error rate and the power. The data were generated according to the order restricted one-way ANOVA model specified in Equation (2.4), $Y_{ij} \sim N(\lambda\mu_i, \tau^{-1})$, with $\tau = 1$ and varying λ. The simulation represented an experiment with $K = 4$ dose levels and followed the experimental design described in Section 6.1. The number of observations per dose level was equal to $n = 3$. The permutation test, introduced in Section 3.2.1.2, was performed using $B = 1,000$ permutations. The null hypothesis was rejected whenever $p_{\text{Bayes}} < 0.05$. The performance of the BVS model was compared with Williams' and Marcus' contrast based MCTs and with the LRT.

A second simulation study was conducted in order to evaluate the findings obtained in Section 5.3 and to explore the dependency of the posterior expected complexity and estimation on the specification of priors. The simulation consists of an experiment with $K = 4$ dose levels with $n = 3$ observations per dose and followed the design described in Section 6.1. The value of $\lambda = 2$ and $\sigma^2 = 1$ were used. In total, $N = 1,000$ data sets were generated for each combination of mean structure and λ.

A third simulation study was conducted in order to explore the impact of noise on the performance of BVS estimation, model selection and complexity. The inference was not study further because of clear robustness against prior configuration and close correspondence to LRT that were demonstrated in the first simulation study (see Section 5.4.1). Analogously to the first two simulation studies, the different configurations of priors were retained in order to compare the impact on model selection. The study design followed the design of the first study, but the data were generated only under model g_5, with $\lambda = 2$ and σ varying from $0.001, \ldots, 5$.

5.4.1 Inference

The aim of first simulation study was to explore the robustness of permutation test against the prior configuration. The LRT and MCTs were compared with the BVS in terms of the control of Type I error and the power of the test. For all the methods, an assumption of a non-decreasing trend was made. The posterior probabilities for models, $\bar{P}(g_r|\text{data})$, were computed for the BVS model, considering all priors listed in Table 5.1. The permutation test was applied in order to test the null hypothesis against an ordered alternative. The null hypothesis was rejected whenever $p_{\text{Bayes}} < 0.05$. Figure 5.5 presents the simulation results for $\lambda = 2$, displaying Type I error rate and power. It can be clearly seen that both the Type I error and the power are not affected by the configuration of the priors.

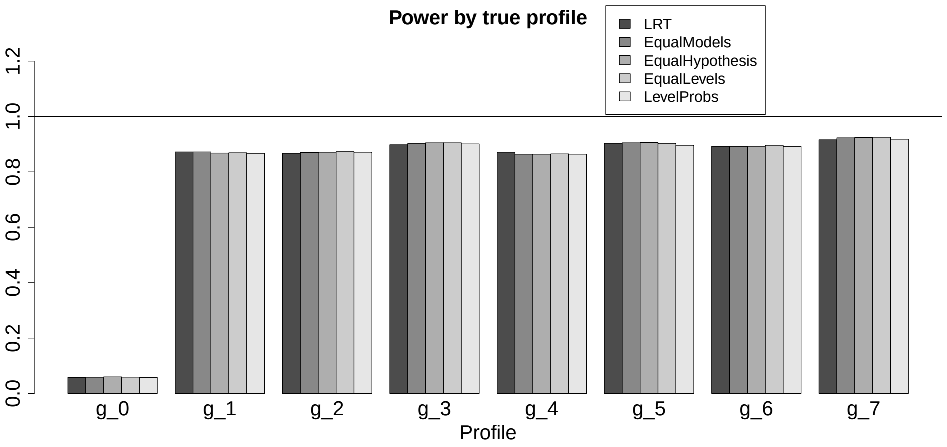


Figure 5.5: *Type I error and power for different prior configurations. Results of the simulation study for $n = 3$ and $K = 4$, with $\lambda = 2$. Each set of bars shows the power of rejecting the null hypothesis, if data were generated under the particular profile $g_1, \ldots, g_7$. In case of g_0, displayed quantity is the Type I error rate. Grey scale distinguishes among different BVS priors: darkest to lightest for E_M, E_H, E_L and L_P, respectively.*

As mentioned before, this results is expected and it is a consequence of conditioning the inference procedure on the priors.

Table 5.2: Results of simulation study for $K = 4$ and $n = 3$. Each row shows proportion of true underlying model being selected as best model according to value of posterior probability. The posterior probability was estimated with BVS model under varying priors and data were generated under the particular profile and λ value. Result of each row is based on mean of 1,000 experiments.

λ	Model	Eq. models	Eq. hypothesis	Eq. levels	Level probs.
	g_0	0.73	0.96	0.90	0.83
2	g_1	0.83	0.70	0.78	0.86
	g_2	0.77	0.65	0.72	0.82
	g_3	0.21	0.21	0.18	0.10
	g_4	0.81	0.70	0.77	0.86
	g_5	0.42	0.40	0.38	0.26
	g_6	0.22	0.22	0.19	0.09
	g_7	0.01	0.01	0.17	0.00
	Mean	0.50	0.48	0.51	0.48

5.4.2 Model selection

The first goal of the second simulation study was to evaluate the sensitivity of the model selection procedure of the BVS to the configuration of the priors. For each simulated data, the model with the highest posterior probability was selected. The results shown in Table 5.2 reveal that, as expected, performance for given model strongly depends on the chosen priors. Naturally, the correct selection rate increases with the higher prior probability assigned to the true underlying model. Analogously, an assignment of high prior probability for the competing models leads to a decrease in the probability of correct selection of true model. Interestingly, there is absence of the overall best method, since average performance across all models for any of the methods is around 0.5. This implies that correct model will be identified in 50% of cases. In the remaining 50%, usually models similar to the true model would be selected.

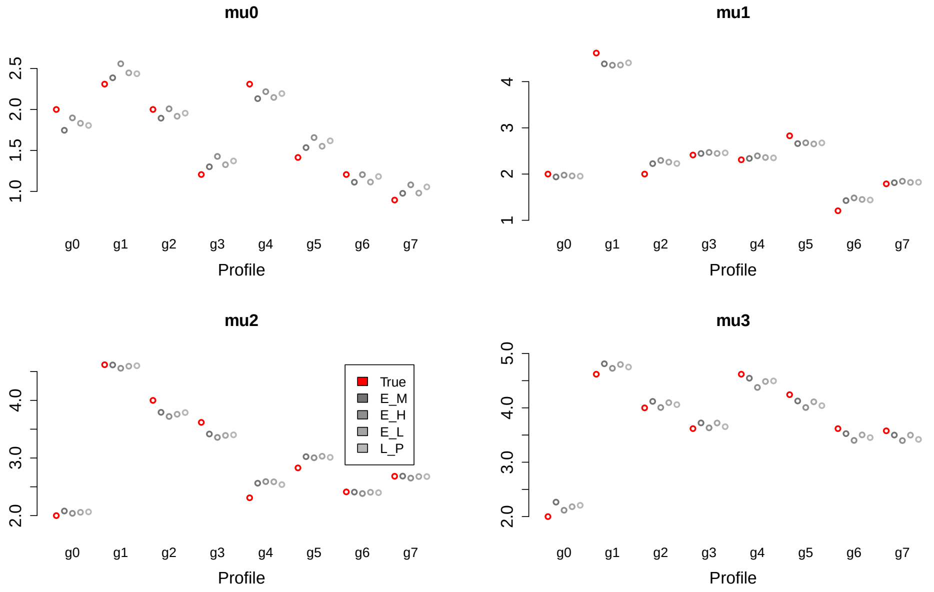


Figure 5.6: *Estimated posterior means for different doses under different prior configuration. Results of the simulation study for $n = 3$ and $K = 4$, with $\lambda = 2$. Each set of points represents estimates for data generated under the profile $g_0, \ldots, g_7$. For each dose level, the leftmost points (red) show the true value of means (according to simulation setting). Grey scale distinguishes among different BVS priors: darkest to lightest for E_M, E_H, E_L and L_P, respectively.*

5.4.3 Estimation

The second goal of the second simulation study was to evaluate the estimation of dose-specific means. The results are evaluated visually and shown in Figure 5.6, where the posterior means based on different prior configurations are compared to true values of $\mu_0, \ldots, \mu_3$ (red point in the left of each dose level). There are differences between the values of estimates, most robust seem to be the estimate of μ_1 and μ_2 for most models. This is expected due to fact that we have more information about the shape in the central part then in the borders of the measurement space. The estimates for the posterior mean of μ_0 and μ_3 are expected to be less precise than estimates for μ_1, μ_2. Although the posterior model probabilities are strongly affected by choice of the priors, the model averaging compensated for this sensitivity and resulted in a relative robustness of dose-specific estimates.

Table 5.3: Expected number of levels for different prior configurations. Results of simulation study for $K = 4$ and $n = 3$, with data generated with given number of levels. In each iteration, the random model with given number of levels was ran (i.e. for two levels one of the models g_1, g_2, g_4). Result of each row is based on mean of 1,000 experiments.

Level	Eq. models	Eq. hypothesis	Eq. levels	Level probs	Isotonic regression
1	1.68	1.26	1.44	1.50	2.09
2	2.32	2.15	2.32	2.16	2.92
3	2.51	2.34	2.56	2.31	3.38
4	2.59	2.45	2.69	2.38	3.60

5.4.4 Posterior complexity

The third goal of the second simulation study was to evaluate the estimation of the posterior expected complexity. In total 1,000 data sets were generated. For a given level, one of the models with corresponding number of levels was randomly selected. For example, for two levels, one of the models g_1, g_2 or g_4 was randomly chosen and a data set was generated according to the chosen model. The results are shown in Table 5.3. Note that there is a relationship between the results presented in Table 5.3 and Table 5.2. High proportion of correct selection implies that the model itself has high posterior probability and complexity pEC should be close to that particular model. For example, for E_H, we have 0.96 probability of selecting g_0 as the correct model. As a result, $\bar{P}(g_0|\text{data})$ is high under this setting and pEC is close to value of one, which is true number of levels under model g_0. In contrast, none of the methods is able to correctly identify model g_7 with high proportion, so pEC is much lower than the true number of levels under g_7 which equals to four.

Note the clear difference between results of the BVS methods and results of isotonic regression. The isotonic regression performs maximum likelihood estimation and ignores any model uncertainty. Consequently, the estimated models tend to have large number of parameters regardless if the true underlying model is complex (relatively to other models in set of candidate models) or if underlying model is simple.

5.4.5 Varying noise

As expected, the correct identification of underlying model and true MED are improved with decrease of the magnitude of the noise in the data, as is shown in Figure 5.7. Note that the performance of the BVS model under different prior configuration is related to the choice of model g_5. The configuration E_L performs consistently worse than E_H and E_M for correct model specification. It is due to high prior probability assigned to model g_7 (see Figure 5.2) that is very similar to model g_5, causing frequent selection of g_7 as the best model. In contrast, E_L performs much better in terms of the MED selection. Due to the the high prior assigned to g_7, the BVS with priors E_L prefers complex models that have the same MED as model g_5 (i.e. the MED is the first dose). Interestingly, L_P seems to work as the best choice for the correct model selection for lower σ, but it performs as the worst for relatively higher levels of noise. The answer can be again found in Figure 5.2: L_P assigns very low prior on model g_7, so there is a low misclassification in the direction of this model. However, with higher level of noise, i.e. larger influence of priors, high probabilities for g_1 and g_4 cause preference of simpler models then g_5.

The most interesting findings are related to the estimation of dose-specific means. In Figure 5.8, we can see how value of μ changes nearly monotonically with increasing σ. While μ_0 and μ_2 increase, μ_1 and μ_3 decrease. This behaviour is related to the shape of the model g_5: increasing effect in first and third dose. With higher level of noise and therefore higher uncertainty about the estimates, the means are shrunk to the overall mean and the null model receives higher posterior probability. This process is demonstrated in Figure 5.9, where whole profiles are shown for few values of σ. With lower level of noise, model g_5 seems to be the clear choice, but with increasing level of noise, the model selection process is becoming more and more uncertain. Additionally, the influence of prior specification increases with higher level of noise, since the amount of information in the data decreases.

The behaviour of posterior expected complexity mirrors behaviour of all other properties. In Figure 5.10 is clearly seen that without the noise, model g_5 is selected and therefore correct number of levels estimated. With increasing σ, the model g_7 is selected in some cases, leading to an increment in posterior expected complexity. Note that with higher level of uncertainty, the null model is selected more often and as a results the posterior mean of the null model increases pushing down estimated posterior expected complexity.

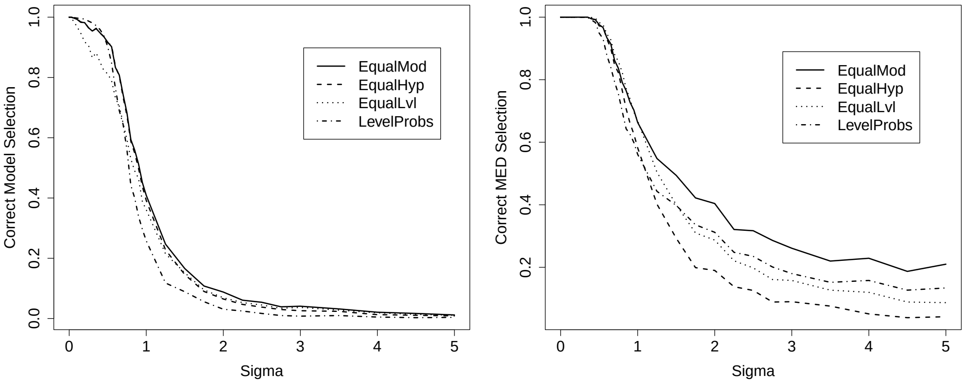


Figure 5.7: *Dependency of correct selection (based on the BVS model) on σ. Left panel: Correct model selection. Right panel: Correct MED selection. Prior configurations: E_M (solid line), E_H (dashed line), E_L (dotted line) and L_P (dash dotted line). Results are based on the simulation study for $n = 3$ and $K = 4$, with $\lambda = 2$ for g_5 and are derived as averages of $N = 1,000$ experiments.*

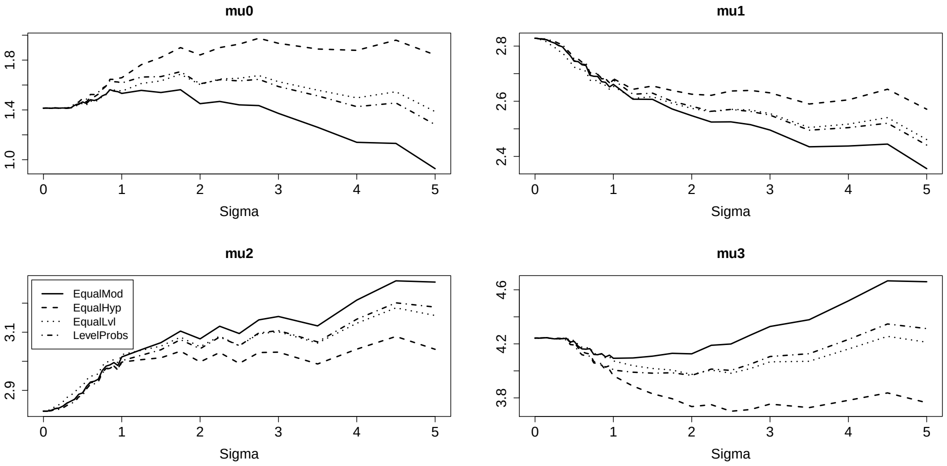


Figure 5.8: *Dependence of the estimates of dose-specific means (based on the BVS model) on σ. The panels shows estimates for dose-specific mean $\mu_0, \ldots, \mu_3$. Prior configurations: E_M (solid line), E_H (dashed line), E_L (dotted line) and L_P (dash dotted line). Results are based on the simulation study for $n = 3$ and $K = 4$, with $\lambda = 2$ for g_5 and are derived as averages of $N = 1,000$ experiments.*

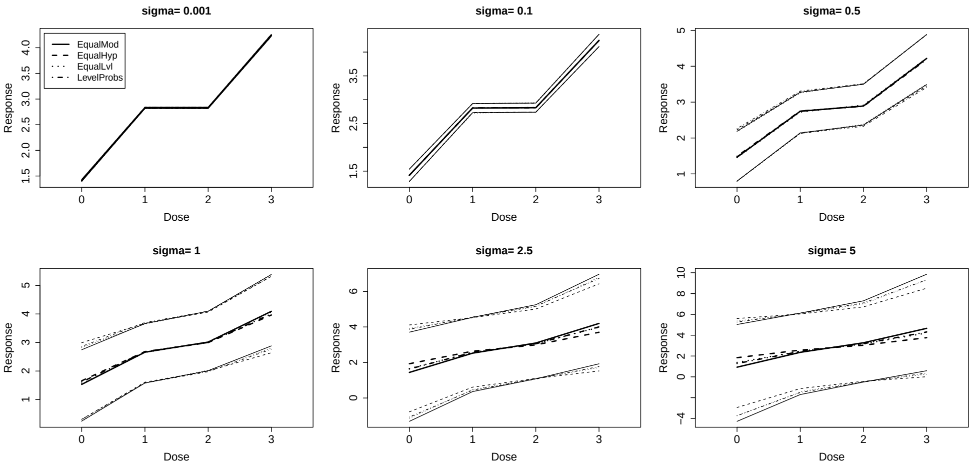


Figure 5.9: *Dependence of the estimates of dose-specific means (based on the BVS model) and their precision on σ. All panels show estimates of whole profile $\mu_0, \ldots, \mu_3$ for a particular choice of $\sigma = 0.001, 0.1, 0.5, 1, 2.5, 5$. The thicker lines in the center represent point estimates, while thinner lines represent 95% credible interval. Prior configurations: E_H (dashed line), E_L (dotted line) and L_P (dash dotted line). Results are based on the simulation study for $n = 3$ and $K = 4$, with $\lambda = 2$ for g_5 and are derived as averages of $N = 1,000$ experiments.*

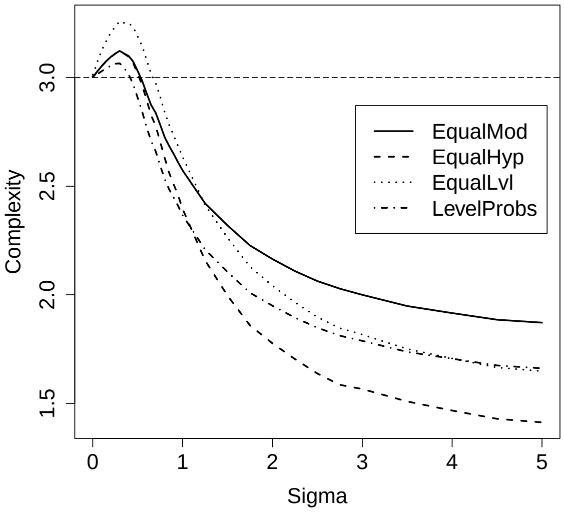


Figure 5.10: *Dependence of posterior expected complexity pEC (based on the BVS model) on σ. Prior configurations: E_M (solid line), E_H (dashed line), E_L (dotted line) and L_P (dash dotted line). The vertical long-dashed line shows true number of levels. Results are based on the simulation study for $n = 3$ and $K = 4$, with $\lambda = 2$ for g_5 and are derived as averages of $N = 1,000$ experiments.*

5.5 Discussion

In this chapter, we presented an investigation about the influence of the specification of non-informative priors on dose-response modelling using the BVS framework. We have introduced four sets of prior configurations that can all be considered non-informative, depending on the primary goals of the analysis. In addition to the estimation, model selection and inference, we focused on expected complexity. Complexity measure was based on ideas arising from information criteria framework.

The simulation study focused on case of $K = 4$ and $n = 3$ only. There was no need for additional settings, because we focus on overall patterns and general robustness, not particular results. Varying K and n would influence the results, but overall patterns shall stay the same. Interesting extension would be to study the influence of the change of the family of prior distributions, e.g. using weakly informative priors (Gelman, 2006).

Inference based on permutation test confirmed its robustness against any choice of priors, since the the permutation procedure is conditioned on the prior configuration. In contrast, model selection and MED specification were both strongly influenced by the specification of the prior model probabilities. Any model selection procedure conducted without strong prior knowledge need to take this fact into account. We have shown that with increasing information in the data (i.e. decreasing noise and/or increasing sample size), the influence of priors was naturally diminishing. Hence, the dependency would be strongest in case of small size experiments. Similarly, the probability of selection of the correct model as the best model decreased with increasing noise. Surprisingly, it exhibit rather stable results for different priors, although the values of the posterior model probabilities themselves changed dramatically across varying priors. Indeed, even when there are severe changes in the values of posterior probabilities, the model with maximal posterior probability could stay the same. Additionally, we expect stable results in the case of low noise level that allows the data to provide enough information. The estimation of the dose-specific means have proven to be even less sensitive to the choice of the priors. The estimates benefit from the fact that if the correct model is not identified, a similar model is often selected instead, providing estimates close to the true underlying means. The Bayesian model averaging approach, considered in this chapter, seems helpful in compensating both for model uncertainty and uncertainty in non-informative priors, leading to stable estimates of dose-specific means. In contrast, posterior expected complexity did not exhibit similar robustness. Its link to the posterior probabilities was much stronger then in case of estimates, because the number of unique means differed across models much more than actual values of dose-specific estimates. Moreover, the models close to the correct model in terms of the profile shape and estimates of dose-specific means are often not close in terms of the number of levels.

In summary, the choice of non-informative priors may influence certain aspects of the analysis, depending on the level of the noise in the data, the amount of observations, the number of dose levels (i.e. the number of candidate models) and the sample size. In case that no prior knowledge can be used and non-informative priors need to be chosen, the absence of unique solution has to be recognized. The potential influence of possible configurations of priors should be evaluated and compared to the amount of information in the data set. If an influence seems strong, then focus on more robust quantities, as dose-specific estimates or hypothesis testing, may be more appropriate. Procedures relying on the posterior model probabilities should be used with extreme caution in such cases. Moreover, even the quantities based on model averaging may be sensitive to the choice of priors, as was demonstrated for the posterior expected complexity.

Chapter 6

Exploring the properties of the Bayesian Variable Selection Modelling Approach: Simulation Studies

Multiple simulation studies were conducted in order to investigate the properties of methods presented in previous chapters. A short description of studies and their results were explained in respective parts of the thesis. This chapter contains more detailed explanation about simulations' settings and provides additional results that were not presented previously. Although each of the simulations studies was designed to evaluate a specific method or property, the core of all simulations' settings was same, as described in Section 6.1. Section 6.2, Section 6.3 and Section 6.4 provides additional results for the methods presented in the Chapter 2, Chapter 3 and Chapter 4, respectively.

6.1 General setting for the simulation studies

The underlying model used to generate the data is the order restricted one-way ANOVA model specified in Equation (2.4), $Y_{ij} \sim N(\lambda\mu_i, \tau^{-1})$, with $\tau = 1$. The value of λ represents different magnitudes of the true dose effect. In the simulations, several values of λ were used, $\lambda = 0, 1, 1.5, 2, 2.5, 3$. Note that $\lambda = 0$ implies that the underlying true model is the null model with no dose effect. Number of observations per dose was set to

$n = 3, 4, 5$.

The configuration for the mean structure $\mu_0, \mu_1, \mu_2, \mu_3$ for $K = 4$ was the same as specified in Marcus (1976), except for the ordering of the models and more λ values. Eight different configurations were used, corresponding to the models $g_0, \ldots, g_7$. The profiles are visualized in Figure 6.1 and presented in Table 6.1. Values for the mean response at each dose levels were multiplied by λ to cover diverse relative differences among the dose levels.

The configuration of the mean structure, $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_4)$, for $K = 5$ was computed following same formulas as for setting of $K = 4$. We defined a vector $\boldsymbol{v}_r$ of non-decreasing integers according to particular model g_r (e.g. for model g_5 it is vector $\boldsymbol{v}_r = (1, 2, 2, 3, 3)$). Then, the final configuration is obtained through the equation

$$\boldsymbol{s}_r = \boldsymbol{v}_r \cdot \frac{\sqrt{K}}{\sqrt{\sum_{j>i}(v_{rj} - v_{ri})^2}}. \tag{6.1}$$

For model g_5, we get $\boldsymbol{s}_r = (1, 2, 2, 3, 3) \cdot \frac{\sqrt{5}}{\sqrt{1+1+4+4+1+1+1+1}}$. In total, sixteen different configurations were used, corresponding to models $g_0, \ldots, g_{15}$. Order restricted relationships are shown in Table 6.2 (for an increasing and decreasing alternatives). The configurations for $K = 5$ are shown in Table 6.3 and in Figure 6.2.

For each of the settings above, 1,000 data sets were generated. The appropriate methods were applied on simulated data set and the results were evaluated. The BVS model was fitted under the assumption of the non-decreasing trend as described in Section 2.4. All frequentists tests were implemented as one sided tests to be consistent with the monotonicity assumed for the BVS and significance level was set to $\alpha = 0.05$.

All the simulations were performed using the package `runjags` (Denwood, In Review) of `R` software (R Core Team, 2014) together with the `JAGS` software (Plummer, 2003). We used for the analyses a Markov Chain Monte Carlo (MCMC) chain of total length 25,000 with a burn-in period of 5,000.

Table 6.1: The configuration for all models was taken from Marcus (1976). The mean structure for $K = 4$ and $\lambda = 1$ (rounded to two decimal places).

Profile	Dose 0	Dose 1	Dose 2	Dose 3
g_1	1.15	2.31	2.31	2.31
g_2	1.00	1.00	2.00	2.00
g_3	0.60	1.21	1.81	1.81
g_4	1.15	1.15	1.15	2.31
g_5	0.71	1.41	1.41	2.12
g_6	0.60	0.60	1.21	1.81
g_7	0.45	0.89	1.34	1.79

Table 6.2: The set of 16 possible monotonic dose-response models for an experiment with five dose levels (including placebo). Denote μ_i the mean response of the dose level. The model g_0 represents the null model of no dose effect.

Model	Up: Mean Structure	Down: Mean Structure
g_0	$\mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4$	$\mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4$
g_1	$\mu_0 < \mu_1 = \mu_2 = \mu_3 = \mu_4$	$\mu_0 > \mu_1 = \mu_2 = \mu_3 = \mu_4$
g_2	$\mu_0 = \mu_1 < \mu_2 = \mu_3 = \mu_4$	$\mu_0 = \mu_1 > \mu_2 = \mu_3 = \mu_4$
g_3	$\mu_0 < \mu_1 < \mu_2 = \mu_3 = \mu_4$	$\mu_0 > \mu_1 > \mu_2 = \mu_3 = \mu_4$
g_4	$\mu_0 = \mu_1 = \mu_2 < \mu_3 = \mu_4$	$\mu_0 = \mu_1 = \mu_2 > \mu_3 = \mu_4$
g_5	$\mu_0 < \mu_1 = \mu_2 < \mu_3 = \mu_4$	$\mu_0 > \mu_1 = \mu_2 > \mu_3 = \mu_4$
g_6	$\mu_0 = \mu_1 < \mu_2 < \mu_3 = \mu_4$	$\mu_0 = \mu_1 > \mu_2 > \mu_3 = \mu_4$
g_7	$\mu_0 < \mu_1 < \mu_2 < \mu_3 = \mu_4$	$\mu_0 > \mu_1 > \mu_2 > \mu_3 = \mu_4$
g_8	$\mu_0 = \mu_1 = \mu_2 = \mu_3 < \mu_4$	$\mu_0 = \mu_1 = \mu_2 = \mu_3 > \mu_4$
g_9	$\mu_0 < \mu_1 = \mu_2 = \mu_3 < \mu_4$	$\mu_0 > \mu_1 = \mu_2 = \mu_3 > \mu_4$
g_{10}	$\mu_0 = \mu_1 < \mu_2 = \mu_3 < \mu_4$	$\mu_0 = \mu_1 > \mu_2 = \mu_3 > \mu_4$
g_{11}	$\mu_0 < \mu_1 < \mu_2 = \mu_3 < \mu_4$	$\mu_0 > \mu_1 > \mu_2 = \mu_3 > \mu_4$
g_{12}	$\mu_0 = \mu_1 = \mu_2 < \mu_3 < \mu_4$	$\mu_0 = \mu_1 = \mu_2 > \mu_3 > \mu_4$
g_{13}	$\mu_0 < \mu_1 = \mu_2 < \mu_3 < \mu_4$	$\mu_0 > \mu_1 = \mu_2 > \mu_3 > \mu_4$
g_{14}	$\mu_0 = \mu_1 < \mu_2 < \mu_3 < \mu_4$	$\mu_0 = \mu_1 > \mu_2 > \mu_3 > \mu_4$
g_{15}	$\mu_0 < \mu_1 < \mu_2 < \mu_3 < \mu_4$	$\mu_0 > \mu_1 > \mu_2 > \mu_3 > \mu_4$

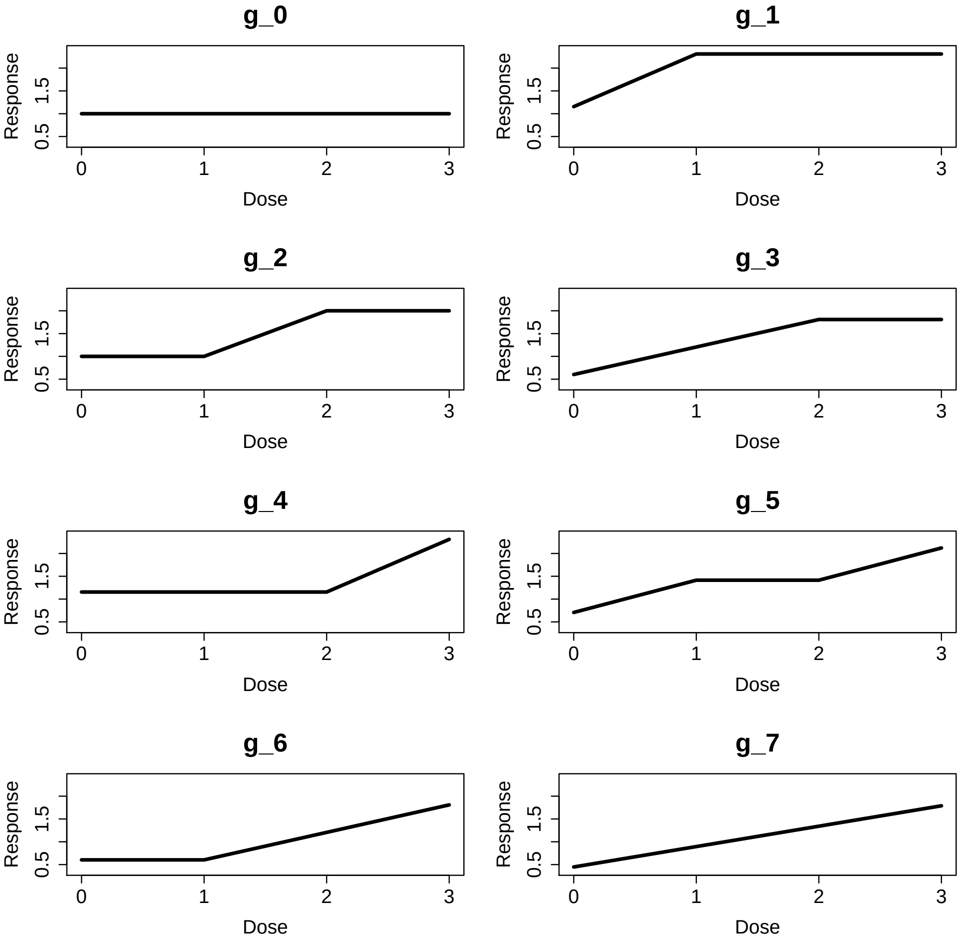


Figure 6.1: *The mean structure for simulation study for $K = 4$ and $\lambda = 1$.*

Table 6.3: The mean structure for simulation study with $K = 5$ and $\lambda = 1$ (rounded to two decimal places).

Profile	Dose 0	Dose 1	Dose 2	Dose 3	Dose 4
g_1	1.12	2.24	2.24	2.24	2.24
g_2	0.91	0.91	1.83	1.83	1.83
g_3	0.56	1.12	1.68	1.68	1.68
g_4	0.91	0.91	0.91	1.83	1.83
g_5	0.60	1.20	1.20	1.79	1.79
g_6	0.56	0.56	1.12	1.68	1.68
g_7	0.40	0.79	1.19	1.58	1.58
g_8	1.12	1.12	1.12	1.12	2.24
g_9	0.71	1.41	1.41	1.41	2.12
g_{10}	0.60	0.60	1.20	1.20	1.79
g_{11}	0.44	0.88	1.32	1.32	1.75
g_{12}	0.56	0.56	0.56	1.12	1.68
g_{13}	0.44	0.88	0.88	1.32	1.75
g_{14}	0.40	0.40	0.79	1.19	1.58
g_{15}	0.32	0.63	0.95	1.26	1.58

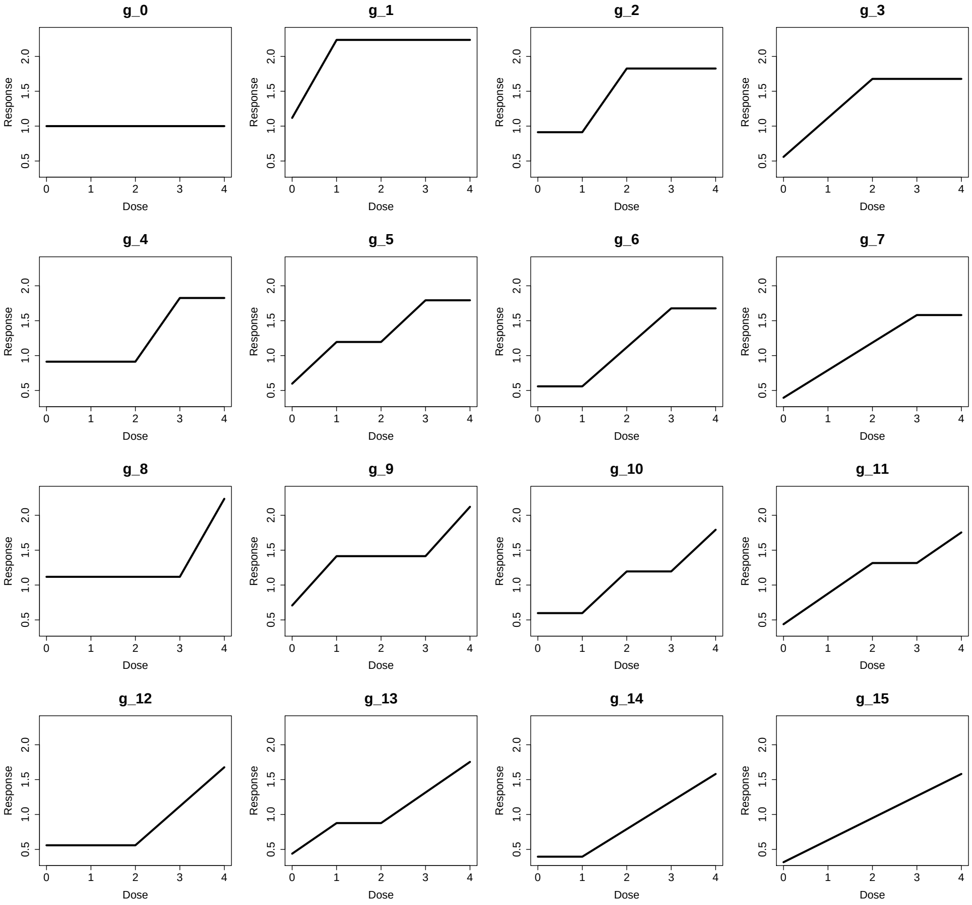


Figure 6.2: *The mean structure for simulation study with $K = 5$ and $\lambda = 1$.*

6.1.1 Model diagnostics

The length of MCMC chains of $L = 20,000$ (with additional 5,000 as burn-in period) is mentioned as sufficient for presented cases of $K = 4,5$ and $n = 3,4$. In this section, we present several diagnostic tools applied on Litter data, discuss their outputs and compare with chains of length $L = 50,000$ (with same burn-in period). Figure 6.3 shows the trace plot of the MCMC chain and the density estimate for the posterior distribution of μ_1 and suggests a good mixing properties for both values of L and indicates that there are no convergence problems. This is supported by the values of the Gelman-Rubin statistic (Gelman and Rubin, 1992) that compares between and within chain variability. A value close to one indicates that the chains were convergent. The Gelman-Rubin statistic in our application was below 1.05 for all parameters for both chain lengths L.

As shown in Figure 6.4, the estimates for the posterior means of model probabilities of both runs (with $L = 20,000$ and $L = 50,000$) are virtually identical.

Prolongation of chain reduces MCMC standard error, i.e. uncertainty due to MCMC simulation, but already for $20,000$ iterations, the error is lower than 4% of the estimated standard error for all the parameters. This indicates that there is no need to focus on longer chains in our framework. In general, when applying the BVS model in more complicated cases, model diagnostic should be performed and MCMC chain's length should be adjusted if necessary.

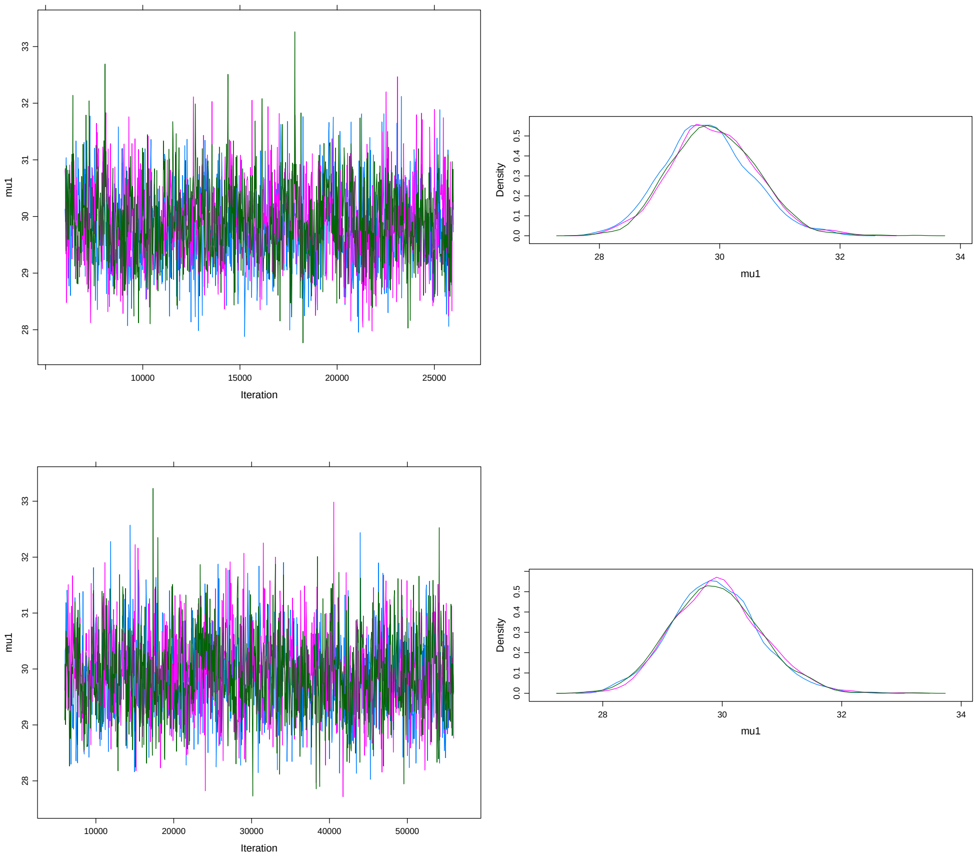


Figure 6.3: *Litter data. Trace plots and density estimates for the posterior distribution of μ_1. The MCMC simulation for the BVS model is based on three chains of length 20,000 (upper panel) and 50,000 (bottom panel). Left figures show mixing of the chains, right figures estimated densities. Each chain is represented by different colour.*

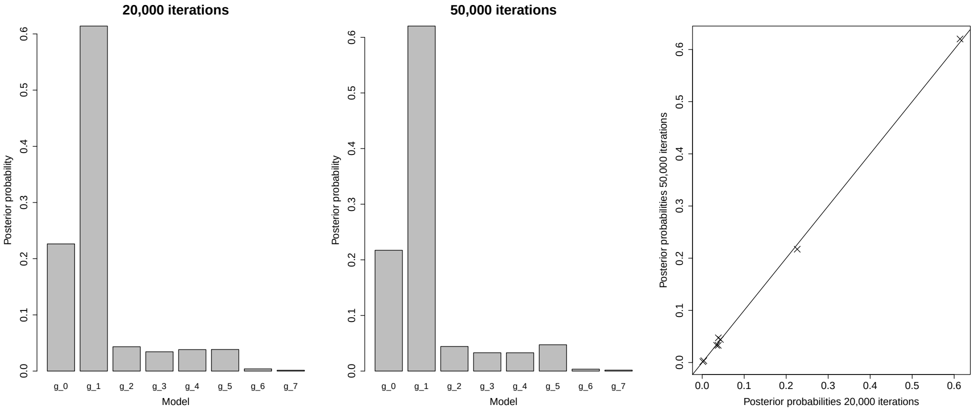


Figure 6.4: *Litter data. Diagnostic plots for estimates of posterior model probabilities* $P(g_r|\textit{data})$. *The MCMC simulation for the BVS model is based on three chains of length 20,000 (left panel) and 50,000 (middle panel). The posterior model probabilities are compared in right panel.*

6.2 Simulation studies: Estimation (Chapter 2)

The simulation study presented in this section was conducted in order to investigate the performance of the BVS model in terms of controlling the Type I error and the power. The simulation settings correspond to the setting described above for an experiment with four and five dose levels. The BVS model was compared with a one-sided LRT and a one-sided MCTs (with both Williams' and Marcus' contrasts).

Figure 6.5, Figure 6.6, Figure 6.7 and Figure 6.8 show comparison of p-values of the frequentist methods and the posterior probabilities of the null model of the BVS for $K = 4$. Both quantities are very different, so we do not expect their correspondence (i.e. points around plotted diagonal line). The figures provides a visualization of the results shown in Table 6.4, Table 6.5 and Table 6.6. The BVS posterior probabilities are higher in absolute values comparing the p-values, hence to achieve similar results (in terms of the power, the Type I error or the number of significant genes identified) the higher threshold than 0.05 has to be used for the BVS. This finding led us to develop the resampling based inference procedure presented in Chapter 3. Note that with an increasing n, the overall power increases and the difference among the methods diminishes in a similar way as with the increasing λ. Visualization of the results is shown in Figure 6.9 for all $n = 3, 4, 5$. The behavior of Type I error was shown in Figure 2.4. The corresponding results for $K = 5$ are provided in Table 6.7, Table 6.8, Table 6.9, Table 6.10, Table 6.11 and Table 6.12. The results of $K = 5$ show the same pattern as was discussed in Chapter 2. Figure 6.10 and Figure 6.11 demonstrate the change in the power when the number of dose levels increase from $K = 4$ to $K = 5$ (for varying λ and n) which corresponds to a change from $1/8$ to $1/16$ for the model prior probabilities, respectively. The results are consistent with the results presented in Chapter 2.

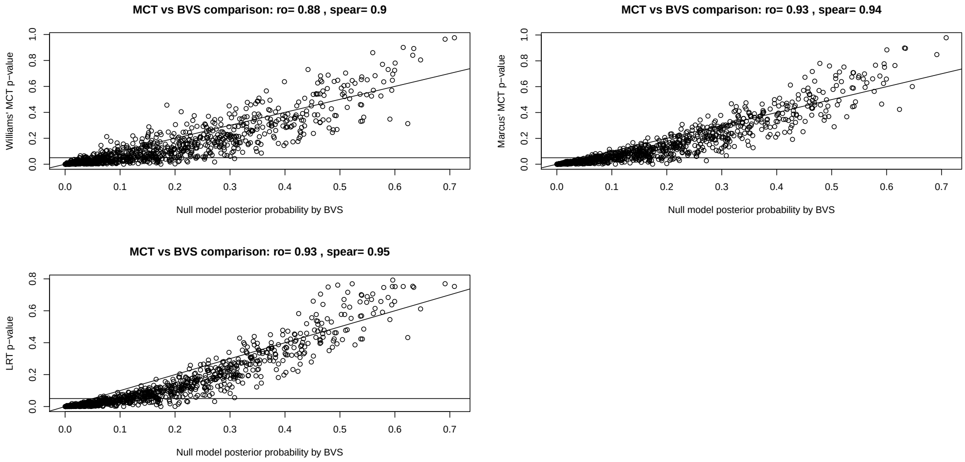


Figure 6.5: *The p-values for the LRT and the MCTs against the posterior probabilities of the null hypothesis obtained by the BVS model. Example for $K = 4$, $\lambda = 1$ and g_7. Top left: LRT vs. BVS. Top right: MCT Williams vs. BVS, Bottom left: MCT Marcus vs. BVS.*

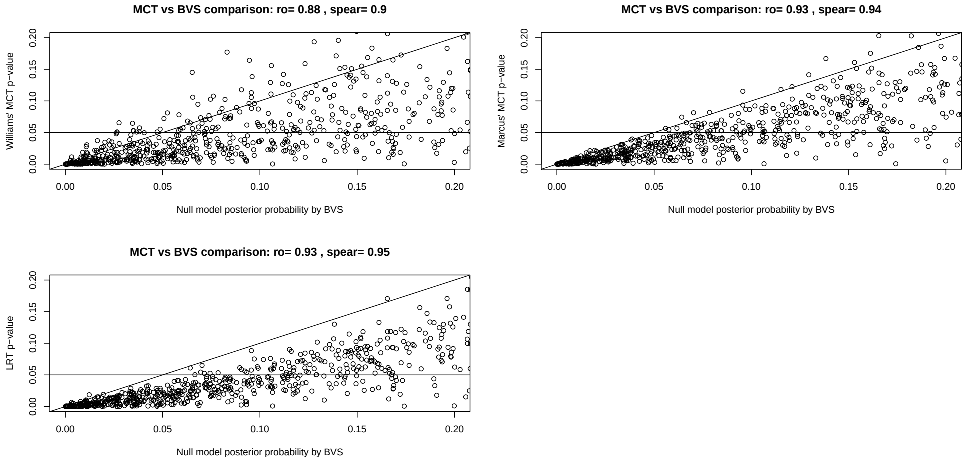


Figure 6.6: *The p-values for the LRT and the MCTs against the posterior probabilities of the null hypothesis obtained by the BVS model. Detail around the zero. Example for $K = 4$, $\lambda = 1$ and g_7. Top left: LRT vs. BVS. Top right: MCT Williams vs. BVS, Bottom left: MCT Marcus vs. BVS.*

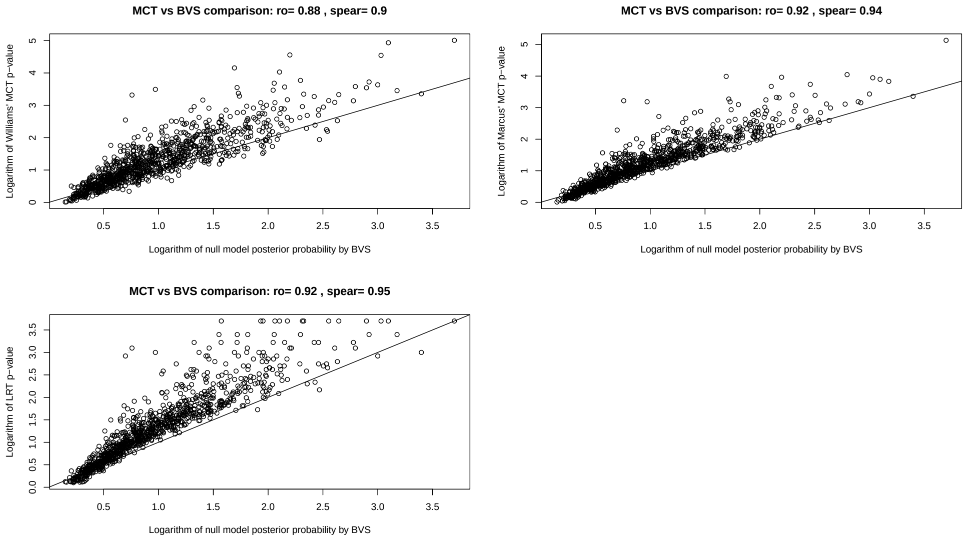


Figure 6.7: *Logarithm of the p-values for the LRT and the MCTs against the logarithm of the posterior probabilities of the null hypothesis obtained by the BVS model. Example for $K = 4$, $\lambda = 1$ and g_7. Top left: LRT vs. BVS. Top right: MCT Williams vs. BVS, Bottom left: MCT Marcus vs. BVS.*

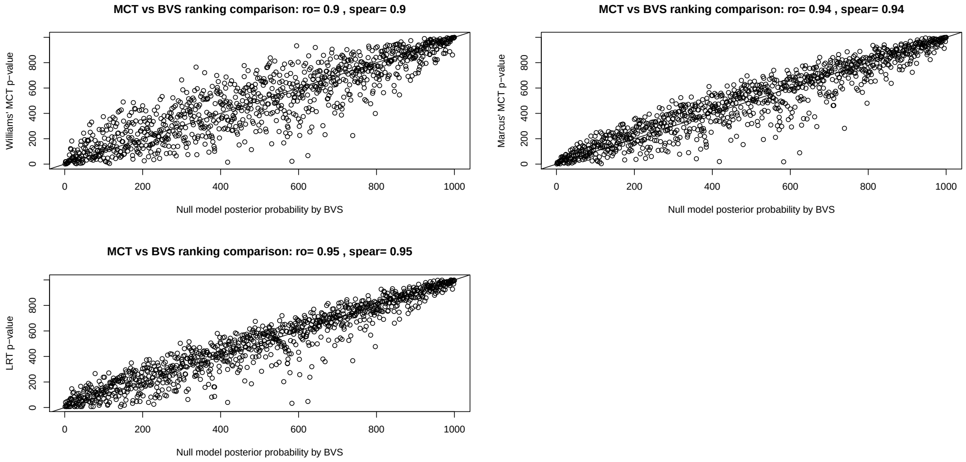


Figure 6.8: *Ranking of the p-values for the LRT and the MCTs against the ranking of the posterior probabilities of the null hypothesis obtained by the BVS model. Example for $K = 4$, $\lambda = 1$ and g_7. Top left: LRT vs. BVS. Top right: MCT Williams vs. BVS, Bottom left: MCT Marcus vs. BVS.*

Table 6.4: Power for the $K = 4$ and $n = 3$. The columns for the BVS show proportion of the posterior probabilities of the null model smaller than $\alpha = 0.05, 0.10, 0.15, 0.35$. Last column represents the estimated probability of the correct model having the highest posterior probability among all the possible models.

			MCT	MCT	BVS	BVS	BVS	BVS	
λ	Profile	LRT	(W)	(M)	0.05	0.10	0.15	0.35	True m.
1	g_1	0.36	0.42	0.34	0.22	0.37	0.49	0.81	0.55
	g_2	0.38	0.31	0.36	0.22	0.37	0.48	0.80	0.46
	g_3	0.40	0.39	0.35	0.22	0.38	0.50	0.83	0.01
	g_4	0.36	0.26	0.35	0.22	0.38	0.50	0.83	0.59
	g_5	0.44	0.42	0.39	0.26	0.41	0.54	0.86	0.07
	g_6	0.41	0.33	0.38	0.22	0.36	0.49	0.82	0.02
	g_7	0.46	0.42	0.41	0.24	0.40	0.52	0.85	0.00
1.5	g_1	0.64	0.71	0.61	0.49	0.67	0.78	0.95	0.73
	g_2	0.68	0.56	0.62	0.51	0.70	0.81	0.96	0.68
	g_3	0.69	0.69	0.67	0.55	0.70	0.81	0.96	0.08
	g_4	0.64	0.51	0.61	0.49	0.65	0.76	0.95	0.75
	g_5	0.70	0.70	0.66	0.54	0.73	0.83	0.97	0.21
	g_6	0.72	0.59	0.67	0.53	0.72	0.82	0.97	0.09
	g_7	0.71	0.67	0.64	0.53	0.72	0.81	0.97	0.00
2	g_1	0.85	0.90	0.85	0.74	0.88	0.93	0.99	0.84
	g_2	0.86	0.73	0.84	0.74	0.88	0.94	0.99	0.78
	g_3	0.89	0.88	0.86	0.81	0.90	0.95	0.99	0.22
	g_4	0.85	0.72	0.82	0.74	0.85	0.92	0.99	0.84
	g_5	0.90	0.91	0.87	0.80	0.92	0.96	1.00	0.42
	g_6	0.90	0.81	0.87	0.80	0.91	0.96	1.00	0.23
	g_7	0.90	0.88	0.87	0.82	0.93	0.97	1.00	0.01
2.5	g_1	0.96	0.98	0.95	0.90	0.97	0.99	1.00	0.86
	g_2	0.96	0.90	0.95	0.90	0.96	0.98	1.00	0.80
	g_3	0.98	0.98	0.96	0.94	0.98	1.00	1.00	0.39
	g_4	0.96	0.90	0.96	0.92	0.97	0.99	1.00	0.87
	g_5	0.98	0.98	0.96	0.94	0.98	0.99	1.00	0.63
	g_6	0.97	0.93	0.96	0.95	0.98	0.99	1.00	0.39
	g_7	0.97	0.98	0.96	0.94	0.98	1.00	1.00	0.04
3	g_1	0.99	1.00	0.99	0.98	0.99	1.00	1.00	0.89
	g_2	0.99	0.98	0.99	0.99	1.00	1.00	1.00	0.84
	g_3	0.99	0.99	0.99	0.98	0.99	1.00	1.00	0.57
	g_4	0.99	0.97	0.99	0.97	0.99	1.00	1.00	0.89
	g_5	1.00	1.00	0.99	0.98	1.00	1.00	1.00	0.79
	g_6	1.00	0.98	0.99	0.99	1.00	1.00	1.00	0.58
	g_7	1.00	0.99	0.99	0.99	1.00	1.00	1.00	0.11

Table 6.5: Power for the $K = 4$ and $n = 4$. The columns for the BVS show proportion of the posterior probabilities of the null model smaller than $\alpha = 0.05, 0.10, 0.15$. Last column represents the estimated probability of the correct model having the highest posterior probability among all candidate models.

			MCT	MCT	BVS	BVS	BVS	
λ	Profile	LRT	(W)	(M)	0.05	0.10	0.15	True m.
1	g_1	0.47	0.54	0.46	0.27	0.41	0.52	0.58
	g_2	0.46	0.38	0.45	0.27	0.40	0.52	0.50
	g_3	0.50	0.52	0.47	0.25	0.41	0.53	0.01
	g_4	0.47	0.36	0.47	0.28	0.42	0.53	0.61
	g_5	0.49	0.50	0.46	0.26	0.40	0.53	0.06
	g_6	0.52	0.42	0.49	0.28	0.43	0.54	0.02
	g_7	0.52	0.50	0.50	0.27	0.41	0.54	0.00
1.5	g_1	0.78	0.83	0.78	0.58	0.75	0.82	0.82
	g_2	0.78	0.67	0.78	0.59	0.74	0.83	0.75
	g_3	0.80	0.81	0.78	0.63	0.77	0.84	0.10
	g_4	0.79	0.67	0.79	0.60	0.75	0.83	0.81
	g_5	0.81	0.82	0.79	0.61	0.77	0.86	0.26
	g_6	0.83	0.74	0.81	0.66	0.80	0.87	0.12
	g_7	0.84	0.82	0.80	0.64	0.80	0.88	0.00
2	g_1	0.95	0.97	0.95	0.86	0.93	0.96	0.86
	g_2	0.95	0.88	0.95	0.84	0.94	0.97	0.84
	g_3	0.98	0.98	0.97	0.90	0.97	0.99	0.29
	g_4	0.96	0.89	0.96	0.88	0.95	0.97	0.86
	g_5	0.97	0.97	0.96	0.88	0.95	0.98	0.52
	g_6	0.96	0.92	0.96	0.88	0.96	0.98	0.27
	g_7	0.97	0.96	0.96	0.90	0.97	0.99	0.02
2.5	g_1	0.99	1.00	0.99	0.98	0.99	1.00	0.91
	g_2	1.00	0.98	1.00	0.97	0.99	1.00	0.86
	g_3	1.00	1.00	1.00	0.99	1.00	1.00	0.50
	g_4	0.99	0.97	0.99	0.98	0.99	1.00	0.90
	g_5	0.99	1.00	0.99	0.97	0.99	1.00	0.74
	g_6	1.00	0.98	0.99	0.97	0.99	1.00	0.49
	g_7	1.00	0.99	0.99	0.98	1.00	1.00	0.06
3	g_1	1.00	1.00	1.00	1.00	1.00	1.00	0.90
	g_2	1.00	1.00	1.00	1.00	1.00	1.00	0.87
	g_3	1.00	1.00	1.00	1.00	1.00	1.00	0.67
	g_4	1.00	1.00	1.00	1.00	1.00	1.00	0.90
	g_5	1.00	1.00	1.00	1.00	1.00	1.00	0.85
	g_6	1.00	0.99	1.00	0.99	1.00	1.00	0.67
	g_7	1.00	1.00	1.00	1.00	1.00	1.00	0.15

Table 6.6: Power for the $K = 4$ and $n = 5$. The columns for the BVS show proportion of the posterior probabilities of the null model smaller than $\alpha = 0.05, 0.10, 0.15$. Last column represents the estimated probability of the correct model having the highest posterior probability among all candidate models.

			MCT	MCT	BVS	BVS	BVS	
λ	Profile	LRT	(W)	(M)	0.05	0.10	0.15	True m.
1	g_1	0.55	0.62	0.54	0.31	0.45	0.55	0.62
	g_2	0.59	0.49	0.59	0.33	0.48	0.59	0.57
	g_3	0.66	0.66	0.64	0.37	0.52	0.65	0.01
	g_4	0.56	0.44	0.57	0.32	0.46	0.56	0.66
	g_5	0.63	0.64	0.59	0.32	0.48	0.61	0.08
	g_6	0.61	0.52	0.60	0.33	0.48	0.59	0.02
	g_7	0.64	0.62	0.61	0.36	0.50	0.62	0.00
1.5	g_1	0.89	0.93	0.89	0.71	0.82	0.90	0.86
	g_2	0.90	0.81	0.90	0.74	0.85	0.91	0.83
	g_3	0.89	0.90	0.89	0.72	0.85	0.90	0.12
	g_4	0.88	0.77	0.88	0.70	0.81	0.87	0.87
	g_5	0.91	0.91	0.89	0.72	0.85	0.92	0.33
	g_6	0.90	0.82	0.90	0.74	0.85	0.91	0.16
	g_7	0.92	0.91	0.90	0.76	0.86	0.92	0.00
2	g_1	0.99	1.00	0.99	0.93	0.97	0.99	0.91
	g_2	0.99	0.95	0.99	0.93	0.97	0.99	0.87
	g_3	0.99	0.99	0.99	0.94	0.98	0.99	0.37
	g_4	0.98	0.94	0.98	0.91	0.96	0.98	0.90
	g_5	0.99	0.99	0.98	0.94	0.97	0.99	0.60
	g_6	0.99	0.97	0.99	0.94	0.98	0.99	0.35
	g_7	1.00	0.99	0.99	0.95	0.99	1.00	0.02
2.5	g_1	1.00	1.00	1.00	0.99	0.99	1.00	0.92
	g_2	1.00	1.00	1.00	0.99	1.00	1.00	0.90
	g_3	1.00	1.00	1.00	0.99	1.00	1.00	0.57
	g_4	1.00	0.99	1.00	0.99	1.00	1.00	0.90
	g_5	1.00	1.00	1.00	0.99	1.00	1.00	0.79
	g_6	1.00	1.00	1.00	1.00	1.00	1.00	0.59
	g_7	1.00	1.00	1.00	1.00	1.00	1.00	0.09
3	g_1	1.00	1.00	1.00	1.00	1.00	1.00	0.93
	g_2	1.00	1.00	1.00	1.00	1.00	1.00	0.90
	g_3	1.00	1.00	1.00	1.00	1.00	1.00	0.79
	g_4	1.00	1.00	1.00	1.00	1.00	1.00	0.93
	g_5	1.00	1.00	1.00	1.00	1.00	1.00	0.90
	g_6	1.00	1.00	1.00	1.00	1.00	1.00	0.76
	g_7	1.00	1.00	1.00	1.00	1.00	1.00	0.23

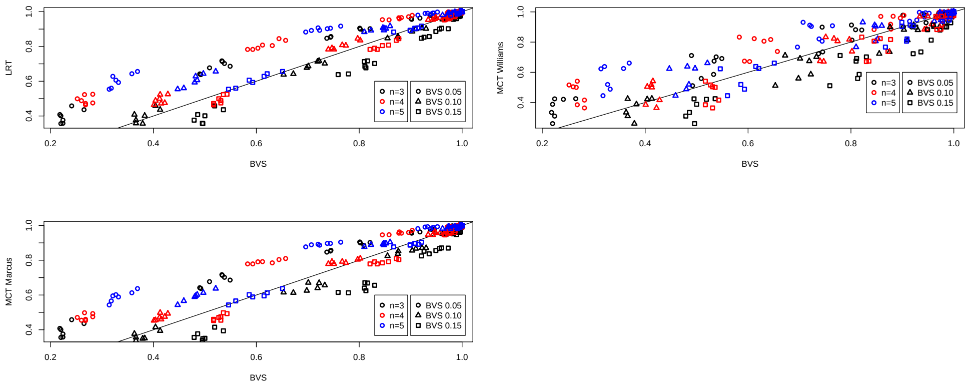


Figure 6.9: *Comparison of the power between the BVS (with varying threshold) and the frequentist tests for* $K = 4$. *Circles represent the results for the threshold* $\alpha = 0.05$, *triangles* $\alpha = 0.10$ *and rectangles* $\alpha = 0.15$. *Black colour is related to the setting of* $n = 3$, *red of* $n = 4$ *and blue of* $n = 5$. *Top left: LRT vs. BVS. Top right: MCT Williams vs. BVS, Bottom left: MCT Marcus vs. BVS.*

Table 6.7: Power for the $K = 5$ and $n = 3$, part 1. The columns for the BVS show proportion of the posterior probabilities of the null model smaller than $\alpha = 0.05, 0.10, 0.15$. Last column represents the estimated probability of the correct model having the highest posterior probability among all candidate models.

			MCT	MCT	BVS	BVS	BVS	
λ	Profile	LRT	(W)	(M)	0.05	0.10	0.15	True m.
1	g_1	0.34	0.44	0.34	0.25	0.40	0.53	0.46
	g_2	0.35	0.29	0.34	0.25	0.39	0.51	0.40
	g_3	0.39	0.40	0.35	0.24	0.39	0.52	0.00
	g_4	0.35	0.25	0.32	0.22	0.39	0.52	0.39
	g_5	0.38	0.41	0.36	0.23	0.42	0.55	0.01
	g_6	0.48	0.37	0.43	0.32	0.49	0.60	0.00
	g_7	0.44	0.41	0.40	0.26	0.41	0.54	0.00
	g_8	0.36	0.28	0.36	0.26	0.41	0.52	0.52
	g_9	0.41	0.42	0.37	0.28	0.43	0.58	0.05
	g_{10}	0.42	0.34	0.36	0.26	0.40	0.54	0.02
	g_{11}	0.42	0.41	0.38	0.26	0.42	0.54	0.00
	g_{12}	0.40	0.28	0.37	0.22	0.39	0.50	0.01
	g_{13}	0.46	0.42	0.41	0.28	0.44	0.54	0.00
	g_{14}	0.48	0.38	0.43	0.27	0.43	0.57	0.00
	g_{15}	0.46	0.40	0.41	0.24	0.40	0.55	0.00
1.5	g_1	0.63	0.73	0.62	0.52	0.70	0.80	0.71
	g_2	0.65	0.54	0.63	0.56	0.71	0.80	0.61
	g_3	0.66	0.67	0.62	0.54	0.72	0.82	0.03
	g_4	0.64	0.48	0.62	0.53	0.71	0.80	0.62
	g_5	0.70	0.72	0.67	0.61	0.77	0.85	0.09
	g_6	0.78	0.65	0.75	0.69	0.84	0.90	0.04
	g_7	0.72	0.70	0.68	0.62	0.77	0.86	0.00
	g_8	0.63	0.50	0.63	0.55	0.70	0.82	0.75
	g_9	0.69	0.72	0.66	0.60	0.76	0.84	0.22
	g_{10}	0.69	0.58	0.65	0.58	0.77	0.86	0.10
	g_{11}	0.72	0.71	0.66	0.61	0.79	0.88	0.00
	g_{12}	0.65	0.50	0.61	0.54	0.70	0.79	0.06
	g_{13}	0.73	0.69	0.69	0.62	0.78	0.87	0.00
	g_{14}	0.74	0.59	0.69	0.61	0.78	0.86	0.00
	g_{15}	0.73	0.67	0.67	0.60	0.75	0.85	0.00
2	g_1	0.83	0.89	0.81	0.76	0.87	0.93	0.82
	g_2	0.86	0.76	0.85	0.81	0.91	0.94	0.76
	g_3	0.88	0.89	0.84	0.82	0.93	0.96	0.13
	g_4	0.87	0.71	0.86	0.82	0.92	0.96	0.73
	g_5	0.89	0.89	0.86	0.83	0.94	0.97	0.25
	g_6	0.94	0.86	0.92	0.91	0.96	0.98	0.13
	g_7	0.93	0.90	0.90	0.88	0.96	0.98	0.00

Table 6.8: Power for the $K = 5$ and $n = 3$, part 2. The columns for the BVS show proportion of the posterior probabilities of the null model smaller than $\alpha = 0.05, 0.10, 0.15$. Last column represents the estimated probability of the correct model having the highest posterior probability among all candidate models.

			MCT	MCT	BVS	BVS	BVS	
λ	Profile	LRT	(W)	(M)	0.05	0.10	0.15	True m.
2	g_8	0.85	0.70	0.83	0.78	0.90	0.94	0.83
	g_9	0.88	0.91	0.86	0.82	0.92	0.95	0.43
	g_{10}	0.89	0.78	0.85	0.82	0.92	0.96	0.27
	g_{11}	0.92	0.90	0.88	0.86	0.95	0.98	0.01
	g_{12}	0.89	0.74	0.86	0.83	0.94	0.97	0.16
	g_{13}	0.90	0.87	0.86	0.84	0.94	0.97	0.01
	g_{14}	0.91	0.80	0.88	0.86	0.95	0.98	0.00
	g_{15}	0.91	0.87	0.86	0.86	0.94	0.97	0.00
2.5	g_1	0.96	0.98	0.95	0.93	0.98	0.99	0.86
	g_2	0.95	0.89	0.94	0.93	0.97	0.99	0.78
	g_3	0.98	0.98	0.97	0.96	0.99	1.00	0.28
	g_4	0.97	0.86	0.97	0.95	0.98	0.99	0.78
	g_5	0.97	0.97	0.96	0.95	0.99	1.00	0.42
	g_6	0.99	0.95	0.99	0.99	1.00	1.00	0.28
	g_7	0.99	0.98	0.97	0.97	1.00	1.00	0.01
	g_8	0.95	0.87	0.94	0.92	0.97	0.99	0.86
	g_9	0.97	0.98	0.96	0.95	0.98	0.99	0.62
	g_{10}	0.98	0.94	0.97	0.97	0.99	1.00	0.45
	g_{11}	0.98	0.98	0.97	0.97	0.99	1.00	0.03
	g_{12}	0.98	0.91	0.97	0.97	0.99	1.00	0.31
	g_{13}	0.98	0.97	0.97	0.97	0.99	1.00	0.05
	g_{14}	0.99	0.94	0.97	0.98	1.00	1.00	0.01
	g_{15}	0.99	0.98	0.97	0.98	0.99	1.00	0.00
3	g_1	0.99	1.00	0.99	0.99	1.00	1.00	0.88
	g_2	0.99	0.97	0.99	0.99	1.00	1.00	0.78
	g_3	1.00	1.00	0.99	0.99	1.00	1.00	0.46
	g_4	1.00	0.94	0.99	0.99	1.00	1.00	0.81
	g_5	0.99	0.99	0.99	0.99	1.00	1.00	0.61
	g_6	1.00	0.99	1.00	1.00	1.00	1.00	0.46
	g_7	1.00	0.99	0.99	1.00	1.00	1.00	0.04
	g_8	0.99	0.95	0.99	0.98	0.99	1.00	0.86
	g_9	1.00	1.00	0.99	0.99	1.00	1.00	0.77
	g_{10}	1.00	0.99	1.00	0.99	1.00	1.00	0.61
	g_{11}	1.00	1.00	1.00	1.00	1.00	1.00	0.10
	g_{12}	1.00	0.97	0.99	0.99	1.00	1.00	0.48
	g_{13}	1.00	0.99	0.99	0.99	1.00	1.00	0.12
	g_{14}	1.00	0.99	1.00	1.00	1.00	1.00	0.04
	g_{15}	1.00	0.99	1.00	1.00	1.00	1.00	0.00

Table 6.9: Power for the $K = 5$ and $n = 4$, part 1. The columns for the BVS show proportion of the posterior probabilities of the null model smaller than $\alpha = 0.05, 0.10, 0.15$. Last column represents the estimated probability of the correct model having the highest posterior probability among all candidate models.

			MCT	MCT	BVS	BVS	BVS	
λ	Profile	LRT	(W)	(M)	0.05	0.10	0.15	True m.
1	g_1	0.44	0.54	0.44	0.28	0.42	0.54	0.52
	g_2	0.48	0.39	0.46	0.30	0.46	0.57	0.45
	g_3	0.49	0.51	0.47	0.29	0.45	0.56	0.00
	g_4	0.47	0.33	0.45	0.30	0.44	0.54	0.49
	g_5	0.56	0.55	0.52	0.34	0.50	0.61	0.02
	g_6	0.61	0.49	0.59	0.41	0.55	0.67	0.01
	g_7	0.57	0.54	0.54	0.33	0.51	0.62	0.00
	g_8	0.44	0.36	0.44	0.30	0.43	0.54	0.58
	g_9	0.52	0.55	0.49	0.31	0.47	0.60	0.08
	g_{10}	0.56	0.45	0.52	0.33	0.49	0.61	0.03
	g_{11}	0.56	0.54	0.53	0.33	0.50	0.61	0.00
	g_{12}	0.49	0.37	0.46	0.28	0.42	0.54	0.01
	g_{13}	0.54	0.50	0.50	0.30	0.47	0.57	0.00
	g_{14}	0.55	0.42	0.52	0.31	0.46	0.56	0.00
	g_{15}	0.53	0.49	0.50	0.28	0.44	0.56	0.00
1.5	g_1	0.78	0.85	0.77	0.62	0.78	0.85	0.80
	g_2	0.79	0.68	0.78	0.65	0.81	0.87	0.69
	g_3	0.82	0.84	0.79	0.66	0.81	0.88	0.04
	g_4	0.77	0.60	0.76	0.65	0.78	0.85	0.69
	g_5	0.83	0.82	0.81	0.68	0.83	0.90	0.10
	g_6	0.90	0.78	0.89	0.80	0.90	0.93	0.06
	g_7	0.85	0.83	0.83	0.72	0.85	0.90	0.00
	g_8	0.77	0.63	0.76	0.63	0.77	0.84	0.81
	g_9	0.83	0.85	0.80	0.67	0.83	0.88	0.24
	g_{10}	0.82	0.72	0.79	0.67	0.81	0.88	0.17
	g_{11}	0.82	0.82	0.78	0.68	0.81	0.88	0.00
	g_{12}	0.81	0.67	0.79	0.68	0.81	0.87	0.06
	g_{13}	0.86	0.84	0.84	0.72	0.85	0.91	0.00
	g_{14}	0.83	0.70	0.81	0.68	0.82	0.88	0.00
	g_{15}	0.83	0.79	0.80	0.67	0.81	0.89	0.00
2	g_1	0.94	0.97	0.95	0.89	0.95	0.97	0.88
	g_2	0.94	0.87	0.94	0.88	0.94	0.97	0.81
	g_3	0.95	0.96	0.94	0.90	0.96	0.97	0.17
	g_4	0.96	0.82	0.94	0.87	0.95	0.97	0.80
	g_5	0.97	0.97	0.96	0.90	0.97	0.99	0.30
	g_6	0.98	0.95	0.98	0.96	0.99	0.99	0.21
	g_7	0.97	0.96	0.96	0.93	0.98	0.99	0.00

Table 6.10: Power for the $K = 5$ and $n = 4$, part 1. The columns for the BVS show proportion of the posterior probabilities of the null model smaller than $\alpha = 0.05, 0.10, 0.15$. Last column represents the estimated probability of the correct model having the highest posterior probability among all candidate models.

			MCT	MCT	BVS	BVS	BVS	
λ	Profile	LRT	(W)	(M)	0.05	0.10	0.15	True m.
2	g_8	0.96	0.86	0.96	0.90	0.95	0.98	0.87
	g_9	0.96	0.97	0.95	0.90	0.95	0.98	0.50
	g_{10}	0.96	0.91	0.96	0.91	0.97	0.98	0.34
	g_{11}	0.97	0.96	0.95	0.92	0.97	0.99	0.02
	g_{12}	0.96	0.87	0.95	0.90	0.96	0.98	0.19
	g_{13}	0.96	0.95	0.95	0.92	0.97	0.99	0.02
	g_{14}	0.97	0.91	0.96	0.94	0.97	0.99	0.00
	g_{15}	0.98	0.96	0.97	0.94	0.98	0.99	0.00
2.5	g_1	0.99	1.00	0.99	0.98	0.99	1.00	0.88
	g_2	0.99	0.97	0.99	0.98	1.00	1.00	0.82
	g_3	0.99	0.99	0.99	0.98	1.00	1.00	0.36
	g_4	0.99	0.95	0.99	0.98	0.99	1.00	0.82
	g_5	1.00	1.00	1.00	0.99	1.00	1.00	0.56
	g_6	1.00	1.00	1.00	1.00	1.00	1.00	0.40
	g_7	1.00	1.00	1.00	0.99	1.00	1.00	0.02
	g_8	0.99	0.96	0.99	0.98	0.99	1.00	0.87
	g_9	0.99	1.00	0.99	0.97	0.99	1.00	0.73
	g_{10}	1.00	0.98	1.00	0.99	1.00	1.00	0.55
	g_{11}	1.00	1.00	1.00	0.99	1.00	1.00	0.06
	g_{12}	0.99	0.96	0.99	0.98	0.99	1.00	0.40
	g_{13}	1.00	0.99	1.00	0.99	1.00	1.00	0.06
	g_{14}	1.00	0.99	1.00	0.99	1.00	1.00	0.02
	g_{15}	1.00	0.99	1.00	0.99	1.00	1.00	0.00
3	g_1	1.00	1.00	1.00	1.00	1.00	1.00	0.89
	g_2	1.00	1.00	1.00	1.00	1.00	1.00	0.85
	g_3	1.00	1.00	1.00	1.00	1.00	1.00	0.59
	g_4	1.00	0.99	1.00	1.00	1.00	1.00	0.86
	g_5	1.00	1.00	1.00	1.00	1.00	1.00	0.71
	g_6	1.00	1.00	1.00	1.00	1.00	1.00	0.57
	g_7	1.00	1.00	1.00	1.00	1.00	1.00	0.07
	g_8	1.00	0.99	1.00	1.00	1.00	1.00	0.89
	g_9	1.00	1.00	1.00	1.00	1.00	1.00	0.85
	g_{10}	1.00	1.00	1.00	1.00	1.00	1.00	0.69
	g_{11}	1.00	1.00	1.00	1.00	1.00	1.00	0.12
	g_{12}	1.00	1.00	1.00	1.00	1.00	1.00	0.58
	g_{13}	1.00	1.00	1.00	1.00	1.00	1.00	0.16
	g_{14}	1.00	1.00	1.00	1.00	1.00	1.00	0.07
	g_{15}	1.00	1.00	1.00	1.00	1.00	1.00	0.00

Table 6.11: Power for the $K = 5$ and $n = 5$, part 1. The columns for the BVS show proportion of the posterior probabilities of the null model smaller than $\alpha = 0.05, 0.10, 0.15$. Last column represents the estimated probability of the correct model having the highest posterior probability among all candidate models.

			MCT	MCT	BVS	BVS	BVS	
λ	Profile	LRT	(W)	(M)	0.05	0.10	0.15	True m.
1	g_1	0.54	0.65	0.53	0.31	0.46	0.58	0.62
	g_2	0.56	0.46	0.55	0.35	0.49	0.59	0.51
	g_3	0.60	0.63	0.58	0.36	0.52	0.62	0.00
	g_4	0.57	0.41	0.57	0.35	0.50	0.60	0.53
	g_5	0.62	0.63	0.60	0.40	0.54	0.64	0.02
	g_6	0.68	0.56	0.66	0.43	0.59	0.70	0.00
	g_7	0.64	0.59	0.61	0.34	0.50	0.64	0.00
	g_8	0.58	0.47	0.58	0.37	0.52	0.61	0.63
	g_9	0.59	0.63	0.56	0.35	0.50	0.61	0.07
	g_{10}	0.59	0.50	0.57	0.34	0.50	0.61	0.03
	g_{11}	0.64	0.62	0.60	0.35	0.53	0.64	0.00
	g_{12}	0.59	0.43	0.58	0.33	0.49	0.59	0.01
	g_{13}	0.63	0.59	0.61	0.39	0.53	0.63	0.00
	g_{14}	0.65	0.51	0.63	0.36	0.52	0.64	0.00
	g_{15}	0.64	0.60	0.60	0.35	0.50	0.61	0.00
1.5	g_1	0.85	0.92	0.86	0.72	0.83	0.89	0.81
	g_2	0.88	0.76	0.87	0.75	0.84	0.90	0.75
	g_3	0.88	0.88	0.86	0.74	0.86	0.91	0.07
	g_4	0.85	0.69	0.84	0.71	0.82	0.88	0.73
	g_5	0.89	0.89	0.87	0.75	0.86	0.91	0.16
	g_6	0.94	0.86	0.94	0.85	0.93	0.96	0.07
	g_7	0.92	0.91	0.91	0.80	0.89	0.94	0.00
	g_8	0.86	0.70	0.86	0.71	0.83	0.87	0.82
	g_9	0.90	0.93	0.89	0.76	0.86	0.92	0.34
	g_{10}	0.91	0.83	0.90	0.79	0.88	0.93	0.18
	g_{11}	0.90	0.90	0.88	0.77	0.88	0.92	0.00
	g_{12}	0.89	0.76	0.88	0.74	0.85	0.92	0.07
	g_{13}	0.90	0.90	0.89	0.78	0.87	0.93	0.00
	g_{14}	0.92	0.82	0.90	0.79	0.89	0.94	0.00
	g_{15}	0.91	0.88	0.89	0.76	0.88	0.93	0.00
2	g_1	0.98	0.99	0.98	0.93	0.97	0.98	0.91
	g_2	0.98	0.94	0.99	0.94	0.98	0.99	0.86
	g_3	0.99	0.99	0.99	0.95	0.99	0.99	0.23
	g_4	0.98	0.91	0.99	0.95	0.98	0.99	0.84
	g_5	0.99	0.99	0.99	0.96	0.98	0.99	0.40
	g_6	1.00	0.98	1.00	0.99	1.00	1.00	0.24
	g_7	0.99	0.99	0.99	0.96	0.99	0.99	0.00

Table 6.12: Power for the $K = 5$ and $n = 5$, part 2. The columns for the BVS show proportion of the posterior probabilities of the null model smaller than $\alpha = 0.05, 0.10, 0.15$. Last column represents the estimated probability of the correct model having the highest posterior probability among all candidate models.

			MCT	MCT	BVS	BVS	BVS	
λ	Profile	LRT	(W)	(M)	0.05	0.10	0.15	True m.
2	g_8	0.99	0.94	0.99	0.95	0.98	0.99	0.90
	g_9	0.98	0.99	0.98	0.94	0.97	0.99	0.63
	g_{10}	0.99	0.97	0.99	0.95	0.98	0.99	0.42
	g_{11}	0.99	0.99	0.99	0.96	0.98	1.00	0.02
	g_{12}	0.99	0.94	0.99	0.95	0.99	0.99	0.26
	g_{13}	0.99	0.99	0.99	0.96	0.99	1.00	0.02
	g_{14}	0.99	0.97	0.99	0.97	0.99	0.99	0.00
	g_{15}	0.99	0.99	0.99	0.97	0.99	0.99	0.00
2.5	g_1	1.00	1.00	1.00	1.00	1.00	1.00	0.90
	g_2	1.00	0.99	1.00	0.99	1.00	1.00	0.88
	g_3	1.00	1.00	1.00	1.00	1.00	1.00	0.44
	g_4	1.00	0.99	1.00	1.00	1.00	1.00	0.85
	g_5	1.00	1.00	1.00	0.99	1.00	1.00	0.60
	g_6	1.00	1.00	1.00	1.00	1.00	1.00	0.50
	g_7	1.00	1.00	1.00	0.99	1.00	1.00	0.02
	g_8	1.00	0.99	1.00	0.99	1.00	1.00	0.90
	g_9	1.00	1.00	1.00	1.00	1.00	1.00	0.81
	g_{10}	1.00	0.99	1.00	1.00	1.00	1.00	0.68
	g_{11}	1.00	1.00	1.00	1.00	1.00	1.00	0.08
	g_{12}	1.00	0.99	1.00	1.00	1.00	1.00	0.49
	g_{13}	1.00	1.00	1.00	1.00	1.00	1.00	0.09
	g_{14}	1.00	1.00	1.00	1.00	1.00	1.00	0.03
	g_{15}	1.00	1.00	1.00	1.00	1.00	1.00	0.00
3	g_1	1.00	1.00	1.00	1.00	1.00	1.00	0.91
	g_2	1.00	1.00	1.00	1.00	1.00	1.00	0.88
	g_3	1.00	1.00	1.00	1.00	1.00	1.00	0.64
	g_4	1.00	1.00	1.00	1.00	1.00	1.00	0.87
	g_5	1.00	1.00	1.00	1.00	1.00	1.00	0.80
	g_6	1.00	1.00	1.00	1.00	1.00	1.00	0.69
	g_7	1.00	1.00	1.00	1.00	1.00	1.00	0.09
	g_8	1.00	1.00	1.00	1.00	1.00	1.00	0.92
	g_9	1.00	1.00	1.00	1.00	1.00	1.00	0.88
	g_{10}	1.00	1.00	1.00	1.00	1.00	1.00	0.81
	g_{11}	1.00	1.00	1.00	1.00	1.00	1.00	0.23
	g_{12}	1.00	1.00	1.00	1.00	1.00	1.00	0.70
	g_{13}	1.00	1.00	1.00	1.00	1.00	1.00	0.24
	g_{14}	1.00	1.00	1.00	1.00	1.00	1.00	0.08
	g_{15}	1.00	1.00	1.00	1.00	1.00	1.00	0.00

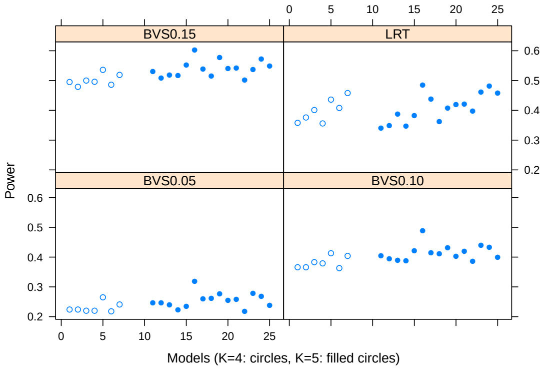


Figure 6.10: *Comparison of the power between $K = 4$ and $K = 5$ for BVS (with varying threshold) and LRT (top right panel). The plot is based on a simulation under $\lambda = 1$ and $n = 3$. The models are ordered arbitrarily, seven models for $K = 4$ on the left (circles) and 15 models for $K = 5$ on the right (filled circles).*

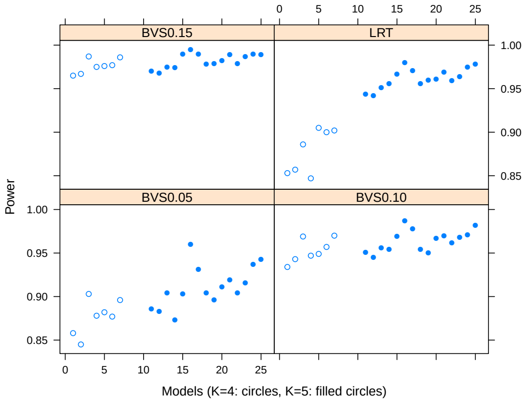


Figure 6.11: *Comparison of the power between $K = 4$ and $K = 5$ for BVS (with varying threshold) and LRT (top right panel). The plot is based on a simulation under $\lambda = 2$ and $n = 4$. The models are ordered arbitrarily, seven models for $K = 4$ on the left (circles) and 15 models for $K = 5$ on the right (filled circles).*

6.3 Simulation studies: Inference (Chapter 3)

The results of the simulation study presented in Section 6.2 indicate that the power obtained for the BVS model and controlling of the Type I error are both dependent on the cut-off point that was used for inference. This cut-off point is not equal to the significance level often used within the frequentist approach. This led us to developed the resampling based inference procedure presented in Chapter 3. The simulation study presented in this chapter was conducted in order to investigate the performance of the resampling based inference procedure for the BVS model in terms of controlling the Type I error and the power.

The simulation settings correspond to the setting described in Section 6.1 and experiment with four and five dose levels was investigated. A sequence of $\lambda = 1, 2, 3$ were used to investigate the magnitude of the differences between the mean response across the doses. Number of observations per dose was equal to $n = 3$ and $n = 4$. The BVS model-based permutation test, one-sided LRT and one-sided MCTs were compared. The permutation test, introduced in Section 3.2.1.2, was performed using $B = 1,000$ permutations. The null hypothesis was rejected whenever $p_{\text{Bayes}} < \alpha$, with $\alpha = 0.05$. The performance of the BVS model was compared with the Williams' and Marcus' contrast based MCT and with the LRT. For all the testing procedures the significance level was set to $\alpha = 0.05$.

Table 6.13 and Table 6.14 present the results of additional settings of simulation study.The results are consistent with the results presented in Section 3. The results for $K = 4$ and $n = 4$ are graphically displayed in Figure 6.12, the results for $K = 5$, $n = 3$ and $\lambda = 1$ are presented in Figure 6.13.

Additional simulation study was conducted that aimed to investigate the Type I error. Therefore, 10^5 separate experiments were simulated, using same mechanism and evaluation as previous studies. The 1,400 experiments were generated under each of the models $g_2, \ldots, g_7$, 1,500 under g_1 and 90,100 experiments were generated under null model g_0. The magnitude parameter was fixed as $\lambda = 2$. The results show the BVS model controls properly for Type I error. Permutation test reached a level of 0.0503, both MCTs 0.0489 and LRT 0.0501. With respect to power, the permutation test is comparable with LRT test (Table 6.15 and Figure 6.14) and results are consistent with previous findings.

Table 6.13: Results for the simulation study with $K = 4$ and $n = 4$. The first row shows the Type I error. Remaining rows show power of rejecting null hypothesis for data that were generated under a particular profile and λ value. Results presented in each row are based on 1,000 experiments.

λ	Profile	MCT (W)	MCT (M)	LRT	BVS
	g_0	0.048	0.042	0.049	0.052
1	g_1	0.568	0.486	0.472	0.465
	g_2	0.416	0.492	0.505	0.525
	g_3	0.553	0.514	0.541	0.545
	g_4	0.361	0.453	0.460	0.492
	g_5	0.569	0.521	0.549	0.543
	g_6	0.442	0.510	0.542	0.574
	g_7	0.546	0.542	0.572	0.586
2	g_1	0.972	0.951	0.950	0.944
	g_2	0.896	0.944	0.955	0.959
	g_3	0.963	0.958	0.964	0.971
	g_4	0.870	0.947	0.951	0.953
	g_5	0.976	0.959	0.969	0.961
	g_6	0.914	0.957	0.966	0.973
	g_7	0.965	0.961	0.973	0.977

Table 6.14: Results for the simulation study with $K = 5$ and $n = 3$. The first row shows the Type I error. Remaining rows show power of rejecting null hypothesis for data that were generated under a particular profile and λ value. Results presented in each row are based on 1,000 experiments.

λ	Profile	MCT (W)	MCT (M)	LRT	BVS
	g_0	0.047	0.048	0.046	0.043
1	g_1	0.439	0.332	0.369	0.356
	g_2	0.309	0.360	0.384	0.394
	g_3	0.417	0.384	0.413	0.398
	g_4	0.275	0.348	0.377	0.404
	g_5	0.410	0.381	0.423	0.420
	g_6	0.371	0.426	0.480	0.511
	g_7	0.416	0.406	0.448	0.458
	g_8	0.265	0.320	0.343	0.380
	g_9	0.424	0.375	0.410	0.401
	g_{10}	0.335	0.369	0.420	0.449
	g_{11}	0.413	0.392	0.436	0.445
	g_{12}	0.295	0.354	0.405	0.460
	g_{13}	0.400	0.388	0.435	0.465
	g_{14}	0.345	0.396	0.453	0.485
	g_{15}	0.396	0.394	0.445	0.476

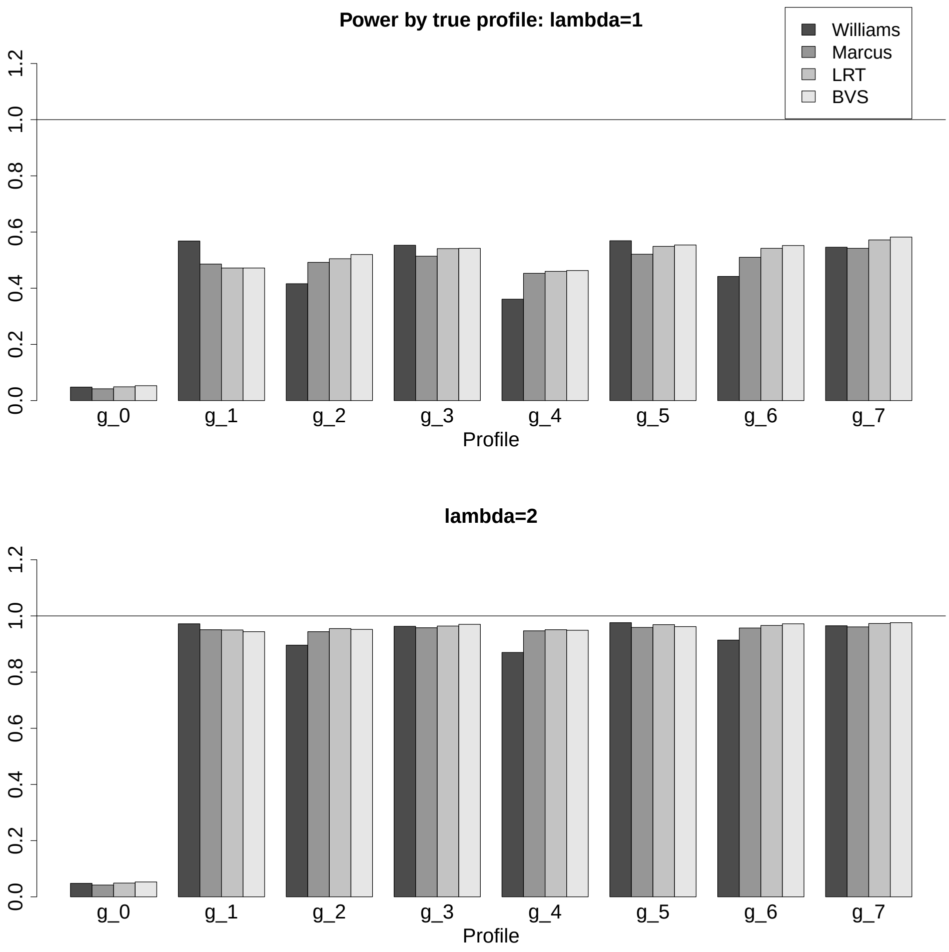


Figure 6.12: *Type I error and power for the simulation study with $n = 4$ and $K = 4$, with $\lambda = 1$ (upper panel) and $\lambda = 2$ (bottom panel). Each set of bars shows power of rejecting null hypothesis, if data were generated under the particular profile $g_1, \ldots, g_7$. In case of g_0, the displayed quantity is the Type I error. Grey scale distinguishes among different tests: darkest for Williams' MCT, then Marcus' MCT, the LRT and brightest for the permutation test. All estimates are based on 1,000 experiments.*

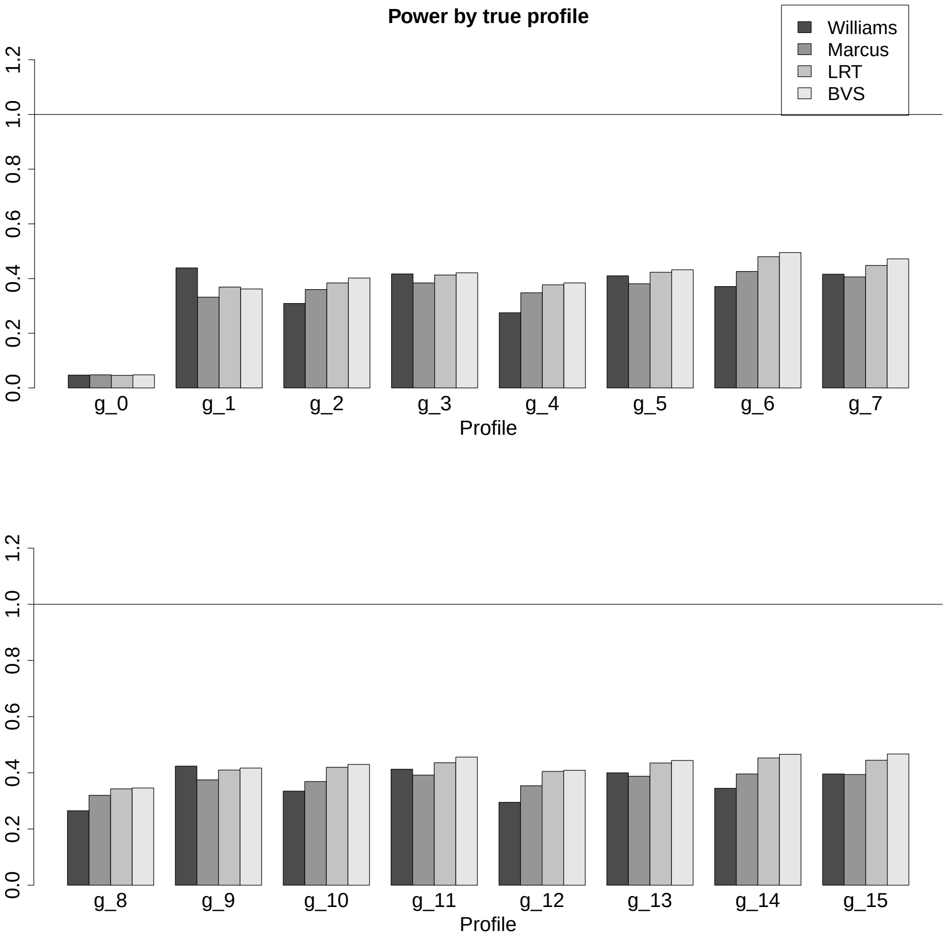


Figure 6.13: *Type I error and power for the simulation study with $n = 3$ and $K = 5$, with $\lambda = 1$. Each set of bars shows power of rejecting null hypothesis, if data were generated under the particular profile $g_1, \ldots, g_15$. In case of g_0, the displayed quantity is the Type I error. Grey scale distinguishes among different tests: darkest for Williams' MCT, then Marcus' MCT, the LRT and brightest for the permutation test. All estimates are based on 1,000 experiments.*

Table 6.15: Results of the second simulation study with $K = 4$ and $n = 3$. First row shows the Type I error. Remaining rows show the power of rejecting the null hypothesis, if data were generated under the particular profile and λ value. All alternative models estimates are based on 1,400 experiments (except g_1 with 1,500 experiments), estimate for g_0 is based on 90,100 experiments.

λ	Profile	MCT (W)	MCT (M)	LRT	BVS
	g_0	0.050	0.050	0.050	0.050
2	g_1	0.898	0.846	0.853	0.853
	g_2	0.767	0.862	0.881	0.877
	g_3	0.886	0.859	0.890	0.894
	g_4	0.735	0.843	0.846	0.854
	g_5	0.890	0.851	0.886	0.885
	g_6	0.789	0.866	0.891	0.909
	g_7	0.879	0.861	0.903	0.911

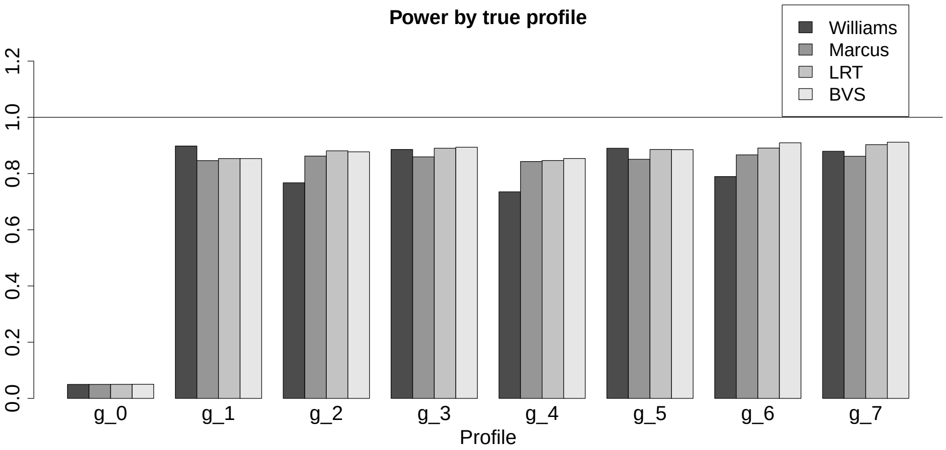


Figure 6.14: *Type I error and power for the second simulation study. Each set of bars shows power of rejecting null hypothesis, if data were generated under the particular profile $g_1, \ldots, g_7$. In case of g_0, the displayed quantity is the Type I error. Grey scale distinguishes among different tests: darkest for Williams' MCT, then Marcus' MCT, the LRT and brightest for the permutation test. All alternative estimates for alternative models are based on 1,400 experiments (except g_1 with 1,500 experiments), estimate for g_0 is based on 90,100 experiments.*

6.4 Simulation studies: Model selection (Chapter 4)

The simulation study presented in this section was conducted in order to explore suitability of various information criteria according to true underlying model. The simulation setting corresponds to an experiment with $K = 4$ dose levels with $n = 3$ observations per dose and followed the design described in Section 6.1. As explained in Chapter 4, not all the models can be fitted for ICs in each simulated data set (when violation of monotonicity in simulated means occurs). Therefore, only suitable models are fitted. in contrast, BVS provides posterior probability for all the models in each simulated data set.

As explained in Chapter 4, the posterior model probabilities can be calculated for all information criteria according to Equation (4.3) as

$$P_{IC}(g_r|\text{data}) = \frac{\exp(-\frac{1}{2}\Delta IC_r)}{\sum_{s=1}^{R}\exp(-\frac{1}{2}\Delta IC_s)}.$$

For the BVS model, $P(g_r|$data) is a part of the quantities estimated by the model (see Section 2.4). The posterior probabilities for MED, $\bar{P}(\text{MED} = i|\text{data})$, are derived by summation of appropriate posterior model probabilities. As in Chapter 4, the methods are evaluated based on two criteria: the identification of correct true underlying model and identification of correct underlying MED.

Table 6.16 shows the rate at which the true underlying model is selected as the best or the second best model. In Table 6.17 and Table 6.18, we can see the probabilities that models would be selected as the best model (or among top two models, respectively), given the true underlying model. These tables show what is the most usual misclassification of the models. The following six tables, Table 6.19, Table 6.20, Table 6.21, Table 6.22, Table 6.23 and Table 6.24, show results of additional settings of simulation study, analogous to the one presented in Chapter 4, but with varying value of n. The results are consistent with the results presented in Chapter 4.

Table 6.16: Comparison of the estimated probability of selection of true model as best or second best model based on 1,000 simulated data sets for BVS, GORIC, AIC and BIC criterion for $K = 4$, $n = 3$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.90	0.82	0.91	0.93
1	g_1	0.77	0.70	0.73	0.74
	g_2	0.64	0.58	0.65	0.66
	g_3	0.23	0.37	0.25	0.20
	g_4	0.76	0.69	0.72	0.72
	g_5	0.32	0.39	0.30	0.25
	g_6	0.21	0.34	0.23	0.19
	g_7	0.00	0.13	0.02	0.01
2	g_1	0.93	0.85	0.91	0.92
	g_2	0.89	0.78	0.87	0.89
	g_3	0.64	0.69	0.64	0.62
	g_4	0.93	0.85	0.89	0.91
	g_5	0.74	0.73	0.72	0.70
	g_6	0.63	0.68	0.63	0.60
	g_7	0.08	0.53	0.21	0.15
3	g_1	0.97	0.88	0.95	0.96
	g_2	0.96	0.82	0.92	0.94
	g_3	0.86	0.86	0.86	0.85
	g_4	0.97	0.88	0.93	0.94
	g_5	0.91	0.89	0.91	0.91
	g_6	0.87	0.85	0.86	0.85
	g_7	0.41	0.84	0.63	0.54

Table 6.17: Selection by the BVS for $K = 4$ and $n = 3$. The probability that a specified model has the highest posterior probability among all candidate models. Rows: The true models. Columns: Selected as the model with the highest posterior probability by BVS. Correct model is shown in bold. Note that the probabilities on the diagonal correspond to the probabilities for the BVS model presented in Table 4.5.

λ	Profile	g_0	g_1	g_2	g_3	g_4	g_5	g_6	g_7
	g_0	**0.73**	0.10	0.08	0.00	0.09	0.00	0.00	0.00
1	g_1	0.21	**0.57**	0.10	0.02	0.07	0.03	0.00	0.00
	g_2	0.22	0.13	**0.46**	0.02	0.14	0.02	0.01	0.00
	g_3	0.19	0.32	0.30	**0.03**	0.10	0.04	0.01	0.00
	g_4	0.22	0.07	0.10	0.00	**0.55**	0.04	0.02	0.00
	g_5	0.18	0.30	0.14	0.02	0.27	**0.08**	0.02	0.00
	g_6	0.20	0.10	0.30	0.01	0.32	0.04	**0.02**	0.00
	g_7	0.17	0.24	0.27	0.03	0.22	0.06	0.02	**0.00**
2	g_1	0.01	**0.83**	0.03	0.06	0.00	0.06	0.00	0.00
	g_2	0.02	0.03	**0.78**	0.06	0.03	0.02	0.06	0.00
	g_3	0.01	0.30	0.34	**0.22**	0.02	0.09	0.02	0.00
	g_4	0.01	0.01	0.02	0.00	**0.82**	0.08	0.06	0.00
	g_5	0.01	0.23	0.05	0.04	0.18	**0.43**	0.05	0.01
	g_6	0.01	0.02	0.34	0.03	0.29	0.08	**0.23**	0.00
	g_7	0.01	0.14	0.28	0.11	0.12	0.22	0.11	**0.01**
3	g_1	0.00	**0.88**	0.00	0.07	0.00	0.05	0.00	0.00
	g_2	0.00	0.00	**0.84**	0.07	0.00	0.00	0.07	0.00
	g_3	0.00	0.15	0.17	**0.59**	0.00	0.06	0.01	0.02
	g_4	0.00	0.00	0.00	0.00	**0.86**	0.08	0.06	0.00
	g_5	0.00	0.06	0.01	0.03	0.05	**0.79**	0.03	0.03
	g_6	0.00	0.00	0.18	0.02	0.14	0.06	**0.57**	0.02
	g_7	0.00	0.03	0.16	0.20	0.02	0.32	0.18	**0.09**

Table 6.18: Selection by the BVS. The probability that a specified model has one of the two highest posterior probabilities among all candidate models. Rows: The true models. Columns: Selected as the model with the highest or the second highest posterior probability by BVS. Correct model is shown in bold. Results for $n = 3$.

λ	Profile	g_0	g_1	g_2	g_3	g_4	g_5	g_6	g_7
	g_0	**0.90**	0.42	0.24	0.02	0.38	0.03	0.02	0.00
1	g_1	0.48	**0.77**	0.19	0.17	0.13	0.23	0.03	0.00
	g_2	0.42	0.26	**0.64**	0.17	0.24	0.07	0.18	0.00
	g_3	0.38	0.52	0.44	**0.23**	0.19	0.15	0.10	0.00
	g_4	0.48	0.15	0.17	0.02	**0.76**	0.24	0.18	0.00
	g_5	0.37	0.46	0.25	0.09	0.42	**0.32**	0.09	0.01
	g_6	0.39	0.20	0.45	0.10	0.49	0.15	**0.21**	0.00
	g_7	0.36	0.40	0.39	0.14	0.35	0.21	0.15	**0.00**
2	g_1	0.09	**0.93**	0.05	0.44	0.02	0.46	0.00	0.01
	g_2	0.08	0.07	**0.89**	0.43	0.08	0.04	0.41	0.01
	g_3	0.04	0.46	0.48	**0.64**	0.05	0.20	0.10	0.04
	g_4	0.10	0.02	0.04	0.00	**0.93**	0.46	0.42	0.01
	g_5	0.04	0.43	0.11	0.10	0.37	**0.74**	0.12	0.09
	g_6	0.04	0.05	0.52	0.09	0.43	0.19	**0.63**	0.04
	g_7	0.03	0.26	0.40	0.31	0.24	0.39	0.28	**0.08**
3	g_1	0.01	**0.97**	0.00	0.50	0.00	0.51	0.00	0.02
	g_2	0.00	0.01	**0.96**	0.52	0.01	0.01	0.47	0.02
	g_3	0.00	0.34	0.42	**0.86**	0.00	0.14	0.05	0.19
	g_4	0.01	0.00	0.00	0.00	**0.97**	0.52	0.48	0.01
	g_5	0.00	0.29	0.02	0.07	0.25	**0.91**	0.08	0.39
	g_6	0.00	0.00	0.45	0.05	0.32	0.13	**0.87**	0.18
	g_7	0.00	0.10	0.30	0.35	0.09	0.44	0.32	**0.41**

Table 6.19: Comparison of the estimated probability of a correct model selection based on 1,000 simulated data sets for BVS, GORIC, AIC and BIC criterion for $K = 4$, $n = 4$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.81	0.64	0.80	0.85
1	g_1	0.63	0.56	0.61	0.58
	g_2	0.51	0.47	0.54	0.50
	g_3	0.03	0.22	0.07	0.03
	g_4	0.56	0.51	0.55	0.51
	g_5	0.10	0.25	0.13	0.08
	g_6	0.03	0.17	0.05	0.02
	g_7	0.00	0.05	0.00	0.00
2	g_1	0.90	0.66	0.84	0.88
	g_2	0.84	0.57	0.79	0.83
	g_3	0.30	0.56	0.40	0.30
	g_4	0.88	0.62	0.80	0.86
	g_5	0.52	0.61	0.59	0.50
	g_6	0.26	0.51	0.35	0.26
	g_7	0.02	0.34	0.07	0.04
3	g_1	0.92	0.66	0.84	0.89
	g_2	0.87	0.57	0.79	0.85
	g_3	0.67	0.73	0.74	0.69
	g_4	0.91	0.62	0.80	0.87
	g_5	0.88	0.70	0.85	0.86
	g_6	0.66	0.68	0.72	0.67
	g_7	0.15	0.72	0.35	0.23

Table 6.20: Comparison of the estimated probability of a correct MED selection based on 1,000 simulated data sets for BVS, GORIC, AIC and BIC criterion for $K = 4$, $n = 4$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.81	0.64	0.80	0.85
1	g_1	0.66	0.79	0.68	0.62
	g_2	0.52	0.56	0.56	0.51
	g_3	0.42	0.59	0.44	0.38
	g_4	0.56	0.51	0.55	0.51
	g_5	0.41	0.60	0.44	0.37
	g_6	0.37	0.44	0.40	0.36
	g_7	0.32	0.48	0.35	0.29
2	g_1	0.98	0.99	0.98	0.97
	g_2	0.88	0.72	0.86	0.88
	g_3	0.66	0.86	0.71	0.64
	g_4	0.88	0.62	0.80	0.86
	g_5	0.75	0.90	0.80	0.74
	g_6	0.60	0.63	0.64	0.63
	g_7	0.50	0.76	0.57	0.49
3	g_1	1.00	1.00	1.00	1.00
	g_2	0.93	0.72	0.87	0.91
	g_3	0.85	0.97	0.90	0.86
	g_4	0.91	0.62	0.80	0.87
	g_5	0.95	0.99	0.96	0.95
	g_6	0.82	0.70	0.81	0.81
	g_7	0.68	0.90	0.78	0.70

Table 6.21: Comparison of the estimated probability of correct model selection based on 1,000 simulated data sets for BVS, GORIC, AIC and BIC criterion for $K = 4$, $n = 5$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.85	0.67	0.81	0.89
1	g_1	0.67	0.62	0.68	0.62
	g_2	0.56	0.51	0.60	0.56
	g_3	0.03	0.25	0.06	0.03
	g_4	0.63	0.58	0.63	0.57
	g_5	0.10	0.29	0.17	0.08
	g_6	0.03	0.23	0.06	0.02
	g_7	0.00	0.07	0.00	0.00
2	g_1	0.92	0.68	0.84	0.90
	g_2	0.88	0.57	0.81	0.88
	g_3	0.35	0.61	0.48	0.34
	g_4	0.91	0.65	0.82	0.89
	g_5	0.59	0.67	0.70	0.58
	g_6	0.34	0.59	0.46	0.33
	g_7	0.02	0.42	0.09	0.04
3	g_1	0.93	0.68	0.84	0.90
	g_2	0.91	0.57	0.81	0.89
	g_3	0.78	0.73	0.83	0.79
	g_4	0.92	0.64	0.82	0.89
	g_5	0.90	0.73	0.88	0.89
	g_6	0.76	0.73	0.79	0.76
	g_7	0.21	0.83	0.45	0.29

Table 6.22: Comparison of the estimated probability of a correct MED selection based on 1,000 simulated data sets for BVS, GORIC, AIC and BIC criterion for $K = 4$, $n = 5$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.85	0.67	0.81	0.89
1	g_1	0.70	0.85	0.75	0.66
	g_2	0.57	0.62	0.62	0.57
	g_3	0.42	0.62	0.44	0.38
	g_4	0.63	0.58	0.63	0.57
	g_5	0.41	0.63	0.47	0.38
	g_6	0.36	0.48	0.42	0.37
	g_7	0.31	0.49	0.36	0.29
2	g_1	0.99	1.00	0.99	0.98
	g_2	0.91	0.75	0.88	0.91
	g_3	0.66	0.90	0.74	0.64
	g_4	0.91	0.65	0.82	0.89
	g_5	0.79	0.94	0.86	0.78
	g_6	0.66	0.69	0.71	0.68
	g_7	0.51	0.77	0.60	0.50
3	g_1	1.00	1.00	1.00	1.00
	g_2	0.94	0.75	0.88	0.93
	g_3	0.89	0.98	0.94	0.90
	g_4	0.92	0.64	0.82	0.89
	g_5	0.96	1.00	0.98	0.97
	g_6	0.88	0.74	0.84	0.87
	g_7	0.69	0.94	0.81	0.72

Table 6.23: Comparison of the estimated probability of correct model selection based on 1,000 simulated data sets for BVS, GORIC, AIC and BIC criterion for $K = 4$, $n = 10$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.90	0.67	0.79	0.91
1	g_1	0.82	0.65	0.80	0.80
	g_2	0.78	0.59	0.78	0.78
	g_3	0.05	0.38	0.18	0.05
	g_4	0.82	0.64	0.80	0.81
	g_5	0.17	0.49	0.35	0.17
	g_6	0.05	0.38	0.16	0.04
	g_7	0.00	0.16	0.01	0.00
2	g_1	0.96	0.66	0.85	0.95
	g_2	0.94	0.61	0.84	0.94
	g_3	0.65	0.74	0.79	0.66
	g_4	0.94	0.65	0.84	0.93
	g_5	0.86	0.70	0.87	0.86
	g_6	0.64	0.75	0.78	0.65
	g_7	0.08	0.76	0.36	0.10
3	g_1	0.96	0.66	0.85	0.95
	g_2	0.95	0.61	0.84	0.94
	g_3	0.96	0.76	0.92	0.96
	g_4	0.95	0.65	0.83	0.93
	g_5	0.97	0.71	0.90	0.96
	g_6	0.95	0.77	0.90	0.94
	g_7	0.57	0.97	0.86	0.64

Table 6.24: Comparison of the estimated probability of a correct MED selection based on 1,000 simulated data sets for BVS, GORIC, AIC and BIC criterion for $K = 4$, $n = 10$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.90	0.67	0.79	0.91
1	g_1	0.85	0.96	0.90	0.83
	g_2	0.79	0.75	0.82	0.79
	g_3	0.46	0.72	0.54	0.42
	g_4	0.82	0.64	0.80	0.81
	g_5	0.48	0.76	0.61	0.46
	g_6	0.46	0.59	0.54	0.49
	g_7	0.30	0.59	0.41	0.29
2	g_1	1.00	1.00	1.00	1.00
	g_2	0.97	0.78	0.92	0.96
	g_3	0.79	0.97	0.90	0.79
	g_4	0.94	0.65	0.84	0.93
	g_5	0.93	1.00	0.97	0.93
	g_6	0.83	0.77	0.86	0.84
	g_7	0.60	0.90	0.74	0.61
3	g_1	1.00	1.00	1.00	1.00
	g_2	0.97	0.78	0.91	0.96
	g_3	0.99	1.00	1.00	0.98
	g_4	0.95	0.65	0.83	0.93
	g_5	1.00	1.00	1.00	1.00
	g_6	0.97	0.77	0.91	0.96
	g_7	0.84	0.99	0.95	0.86

Part II

Microarray Experiments in Toxicogenomics

Chapter 7

Prediction of Human Data Using Rat Data in Japanese Toxicogenomics Project

7.1 Introduction

7.1.1 Toxicogenomics

Pharmaceutical companies are facing urgent needs to increase their lead compound and clinical candidate portfolios and satisfy market demands for continued innovation and revenue growth (Davidov *et al.*, 2003). However, in the last years, relatively small number of drugs are being approved, while research expenses are increasing, patents are expiring, and both governments and health insurance companies are pushing for low-cost medications (Scannell *et al.*, 2012). Moreover, 20-40% of novel drug candidates fail because of safety issues (Arrowsmith, 2011 and Enayetallah *et al.*, 2013), increasing the costs of bringing new drugs to the market (Paul *et al.*, 2010). A significant part of such costs could be prevented if undesirable toxic effects of a potential drug would be predicted in earlier stages of the drug development process (Food and Drug Administration, 2004). Integrating transcriptomics in drug development pipelines is being increasingly considered for early detection of potential safety issues during preclinical development and toxicology studies (Bajorath, 2001, Fanton *et al.*, 2006, Baum *et al.*, 2010 and Amaratunga *et al.*, 2014). Such an approach has proven useful both in toxicology (Pognan, 2007, Afshari *et al.*, 2011) and carcinogenicity studies (Nie *et al.*, 2006, Ellinger-Ziegelbauer *et al.*, 2008).

The approach can be viewed from the perspective of translational research. Translation between rat and human data is an important topic (McGonigle and Ruggeri, 2014), due to high costs and ethical considerations of clinical experiments in humans (Hobin *et al.*, 2012). Gaining strong scientific knowledge in animal models would prevent most risks. Translational research gets attention in all medical fields (e.g. Andrews, 2013, Mestas and Hughes, 2004) and genes are a valuable tool in revealing connections across species (e.g. Seok *et al.*, 2013, Rye *et al.*, 2011).

This chapter and following Chapter 8 focus on translational research in context of toxicogenomics. The Japanese Toxicogenomics Project (TGP) data set was introduced in Section 1.2, where the elaborate description of the data set is given. In summary, the data consists of human *in vitro* experiment and rat *in vitro* and *in vivo* experiments. For both *in vitro* and *in vivo*, dose-response experiments at different time points were conducted. In total, gene expression was measured for 131 compounds.

The TGP data allows to explore two directions of translation important within drug development process: translation across species and across platforms. The former one is import in safety studies to prevent avoidable toxicity in humans. To proceed from animal research to treating patients, we have to assume that animal model predicts toxicity in humans sufficiently well. We focus on this aspect in this chapter, while the translation across platforms is addressed in Chapter 8.

7.1.2 Prediction of human *in vitro* data

The main topic that we address in this chapter is related to the prediction of drug-induced liver injury (DILI) in humans using rat *in vivo* data (henceforth referred to as rat data). The analysis can be viewed from the perspective of translational research. Our aim is to explore the connection between humans and rats in terms of translatability of gene expression. Particularly, our goal is to investigate the effect of a compound on human *in vitro* toxicogenomic data (henceforth referred to as human data) using rat data. Therefore, our method enables identification of genes with toxic effects in rats translatable to humans. Successful prediction of a compound being toxic during rat experiments could reduce the failure rate of efficacious compounds during the expensive phase III trials.

The core part of the rat data set is gene expression level information across multiple compounds at multiple time points and dose levels. We focus on genes that are orthologous for rats and humans. Most of these genes are already annotated by biological processes or diseases (e.g. Ashburner *et al.*, 2000, Lamb *et al.*, 2006). The analysis presented in this chapter explores common dose-response pathways between rat and human genome using gene expression. Identifying a subset of genes that show similar dose-response gene expression profiles in rats and humans would support the translation of gene expression from rat *in vivo* experiments to human experiments. As in the case of DILI, this would enable prediction of compounds' toxicity in humans using rat *in vivo* experiments. The discovery of such genes would create knowledge about underlying mechanisms and connection between species which would significantly improve how rat toxicology is used as a model for human toxicology in the later stages of drug development.

The Translatability data described in Section 1.2.2.1 are used for the analysis of this chapter. Methodology is introduced in Section 7.2 and analysis is conducted in Section 7.3. The results are put in context in Section 7.4 that concludes the chapter.

7.2 Methods

7.2.1 Exploratory analysis: Analysis of variance approach

For the exploratory analysis, a gene specific linear model with dose and time as covariates is used. Interaction between covariates is also included. Let Y_{ijk} denotes the gene expression level for the ith compound ($i = 1, \ldots, 93$), jth gene ($j = 1, \ldots, 4359$) and kth observation ($k = 1, \ldots, 48$ or 36) based on time-dose combinations. To test possible dose effect, time effect and as well as their interaction, a two-way analysis of variance (ANOVA) model is used:

$$Y_{ijk} = \boldsymbol{\alpha}_{0ij} + \boldsymbol{\beta}_{Dij}\text{Dose}_{ijk} + \boldsymbol{\beta}_{Tij}\text{Time}_{ijk} + \boldsymbol{\gamma}_{ij}\text{Dose}_{ijk} \cdot \text{Time}_{ijk} + \varepsilon_{ijk}.$$

Parameters $\boldsymbol{\alpha}_0, \boldsymbol{\beta}_D, \boldsymbol{\beta}_T, \boldsymbol{\gamma}$ are gene (within compound) specific and the measurement error ε_{ijk} is considered to follow a Gaussian distribution $\varepsilon_{ijk} \sim N(0, \sigma_{ij}^2)$. The parameter vectors $\boldsymbol{\beta}_{Dij}, \boldsymbol{\beta}_{Tij}, \boldsymbol{\gamma}_{ij}$ represent the dose, time and interaction effects. In practice, each vector represents levels of explanatory variables, e.g. $\boldsymbol{\beta}_{Dij} = (\beta_{DijCONTROL}, \beta_{DijDOSE1}, \beta_{DijDOSE2}, \beta_{DijDOSE3})$. Note that the two-way ANOVA model specified above is fitted as a gene specific model within each compound. Testing if the parameters differ from null gives us an indication if the gene is differentially expressed for a given compound, or not. However, gene specific omnibus test based on F-distribution can also be used to test if there is any significant effect at all.

Whatever test is used, multiplicity adjustment have to be applied due to extensive number of tests performed (4,359 per compound). Correction for multiplicity was applied within each compound. In general, either Family Wise Error Rate (FWER, Hochberg and Tamhane, 1987) or False Discovery Rate (FDR, Benjamini and Hochberg, 1995) can be used. Controlling FWER translates into level of certainty that there is no false positive finding among all our findings, but controlling FDR assumes there is at least one false positive finding while controlling for proportion of false positive amongst all findings. Hence, FWER is a more conservative method than FDR. In our analysis, we apply Bonferroni method to control FWER to prevent false positives entering later stages of the analysis.

The whole procedure is conducted for both rat *in vivo* data and human data. Only those genes that are significant (according to test we choose) for both humans and rats are kept for further analysis. The resulting lists of significant genes are compared across compounds to identify genes that are significantly expressed in multiple compounds. Indicators of significance of a particular gene can be compared with DILI status of compounds to find out if the genes' appearance is connected with potential danger for the liver. In general, any information about compounds can be used in this stage and can be compared with indicator of genes' significance. For example, pathological data available in the study can be used, as well as information about compound chemical structure or grouping of compounds based on their phenotypic effect.

7.2.2 Main data analysis: Trend analysis approach

A trend analysis is a common analysis in toxicology. The aim of such analysis is to identify a subset of genes for which a monotone relationship with an increasing dose of a compound can be detected (Lin *et al.*, 2012b). Such an assumption of monotonicity allows us to gain power and it is scientifically reasonable. For toxicological studies, this assumption is typically used, since toxic effect usually gets stronger with increasing dose. Monotone means are computed for each gene using isotonic regression method (Barlow *et al.*, 1972, Robertson *et al.*, 1988, Shkedy *et al.*, 2012a). Isotonic regression pools together the means that violate assumption of monotonicity and makes these means equal. Figure 7.1 shows examples of the isotonic means $\boldsymbol{\mu} = (\mu_0, \mu_1, \mu_2, \mu_3)$ for an experiment with control dose and three active dose levels.

Hence, within the second modelling approach the null hypothesis of no dose effect is tested against an ordered alternative in the following way:

$$\begin{aligned} &H_0 : \mu_0 = \mu_1 = \mu_2 = \ldots = \mu_{K-1}, \\ &\qquad\qquad\qquad \text{vs} \\ &H^{up} : \mu_0 \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_{K-1}, \\ &\qquad\qquad\qquad \text{or} \\ &H^{dn} : \mu_0 \geq \mu_1 \geq \mu_2 \geq \ldots \geq \mu_{K-1}, \end{aligned}$$

with at least one inequality strict. We start with simple ANOVA model:

$$Y_{ijlk} = \mu_{0ijl} + \boldsymbol{\delta}_{ijl}\text{Dose}_{ijlk} + \varepsilon_{ijlk},$$

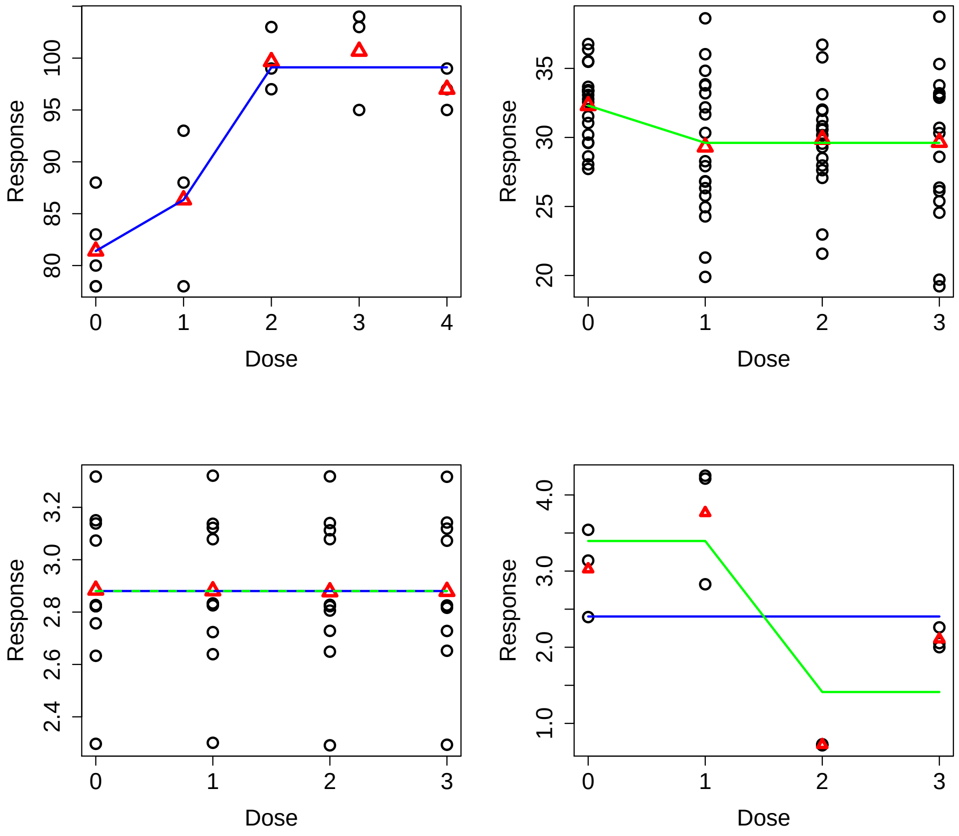


Figure 7.1: *Examples of isotonic regression. Red triangles represent sample means and blue (and green) lines resulting isotonic means, under either upward or downward monotone assumption.*

where i represents compound, j gene, l specific time point and k observation within each time point (within gene, within compound). The vector of parameters $\boldsymbol{\delta}_{ijl} = (\delta_{1ijl}, \delta_{2ijl}, \delta_{3ijl})$ represents the change of the mean in particular dose (compared to control dose) and parameters are either non-negative or non-positive (according to direction of monotonicity assumption). The measurement error follows a Gaussian distribution, $\varepsilon_{ijlk} \sim N(0, \sigma^2_{ijl})$. An advantage of the model is absence of any parametric assumption on dose-response relationship shape. Dose-specific means are modeled separately, connected through values of $\boldsymbol{\delta}_{ijl}$.

The analysis is done per compound and per time point (and separately for human and rat). A multiple contrast test with Marcus' monotone contrast (MCT, Mukerjee *et al.*, 1987) is used to identify significant genes. The MCT is designed to cover space of alternative models while using as few tests as possible (and so keeping power as high as possible). It comprises of several single contrast tests, while different combination of contrasts can also be used. We follow implementation arising from Marcus' test statistics (Marcus, 1976) proposed for MCT by Bretz (2006). Multiplicity adjustment is conducted within each compound and time point combination using FWER approach (with Bonferroni correction) within a gene and FDR adjustment across the genes (Lin *et al.*, 2012b).

Finally, for each compound and time point combination, we create lists of genes that show significant dose-response relationship. The time points with highest rate of significant genes (if such exist) are identified and we focus on them. Then, genes are listed that show significant dose-response relationship for such time points simultaneously in both rats and humans. For a particular gene on the resulting list, isotonic means at all doses are estimated and their values are compared between humans and rats. Hence, we can identify such genes in rats that can be used in order to predict the gene expression level in humans.

7.3 Results

7.3.1 Analysis of variance

Figure 7.2 shows the number of genes with significant interaction effect in both rats and humans and reveals a heterogeneous pattern across compounds. For example, for the compound sulindac there are 201 genes with significant interaction effect in both rats and humans while for the compound perhexiline there is only one gene in common. In total, only 54 compounds had at least one significant gene and only 10 compounds had more than 25 significant genes on the list. An example of one significant gene is shown in Figure 7.3. There exists a small set of genes that are significant in both rat and humans data consistently across subsets of compounds, even in case of strict multiplicity corrections. For the results presented in this chapter we applied Bonferroni correction at significance level of 10%. The subset of compounds, identified through common significant genes, consists of DILI related compounds only (if we convert the DILI status into binary variable, by pooling together "most concern" and "less concern" categories). Hence, the significance of the identified genes in rat *in vivo* could emphasize possible danger of DILI in humans. These genes are typically connected with the liver processes. Table 7.1 shows one of these genes, noted *ASF1A* (originally *Asf1a* in rat and *ASF1A* in human), that is significant for multiple compounds with DILI concern and not for any compound without DILI concern. Other genes from the identified set, *FABP1*, *MCM4*, *SMC2*, *TXNRD1*, show very similar behavior.

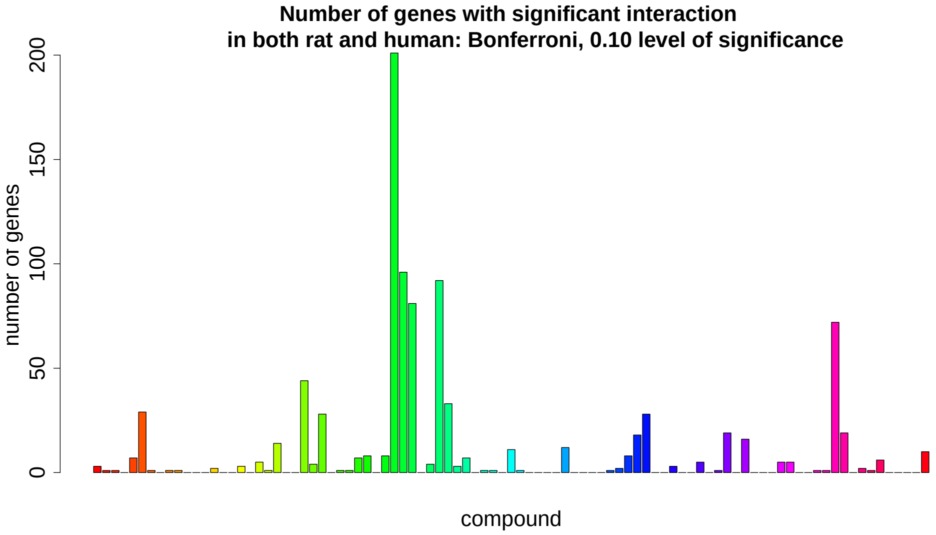


Figure 7.2: *Number of genes with significant interaction in two-way ANOVA model, for both rat and human. The p-values are adjusted using Bonferroni's method on significance level of* 10%.

Table 7.1: Relationship between DILI concern status and simultaneous significance of interaction for both rat and human data for gene *ASF1A*.

	no DILI concern	some DILI concern
non-significant interaction	8	62
significant interaction	0	23

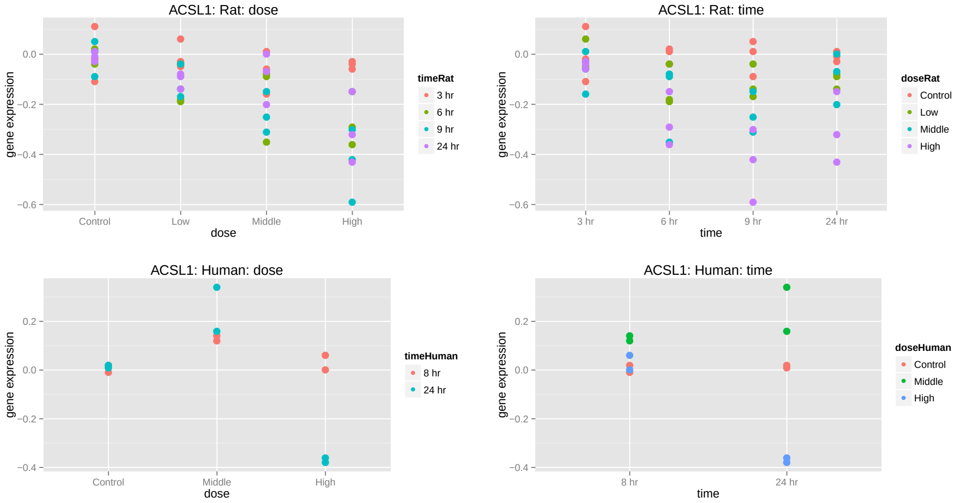


Figure 7.3: *Example of gene with significant interaction in two-way ANOVA model, for both rat and human. Compound omeprazole and gene* Acsl1 *in rat, respectively* ACSL1 *in human.*

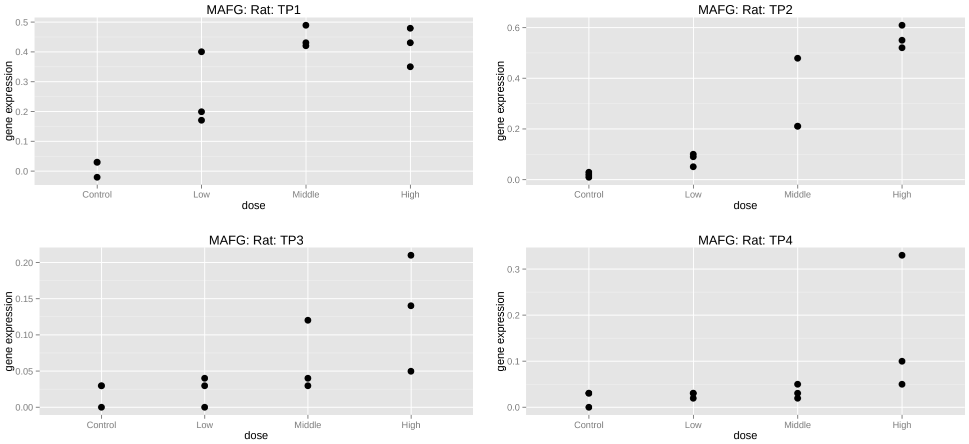


Figure 7.4: *Example of gene with monotone dose-response profile for all time points in rat. Compound omeprazole and gene Mafg in rat (MAFG in human).*

7.3.2 Trend analysis

As mentioned in the previous section, the second analysis consists of trend analysis per time point. An example of gene complying with monotonicity assumption is shown in Figure 7.4. Our aim in this section is to predict dose effect in humans using dose effect in rat *in vivo*. All tests are based on MCT and p-values are adjusted using Bonferroni correction using significance level of 10%.

At the first stage of the analysis, we identify, in the rat, the time point with the strongest signal. Figure 7.5 presents the number of genes with significant dose-response relationship per time point. It clearly shows that there are much more significant genes in the last time point, both for rats and humans, than in any other time point. Hence, for the remainder of this section, the dose effects in rats at the last time point are used for prediction. Figure 7.6 reveals that the number of significant genes in rats does not correspond with the number of significant genes in humans. For several compounds, there are no genes significant both in rats and humans. Hence, we focus on two gene sets: (1) genes significant in rats and (2) genes significant both in rats and humans.

The dose effect in both rats and humans were estimated using isotonic regression. Only 91 compounds having high dose were considered for the analysis and we used the change in isotonic means of rat (from the last to the first, i.e. control, dose level) in order to predict the change in isotonic means of human. The example of resulting gene for the compound omeprazole is presented in Figure 7.7. We can see one of the genes where the translatability of rat data into human data is apparent. The mean at high dose for the rat

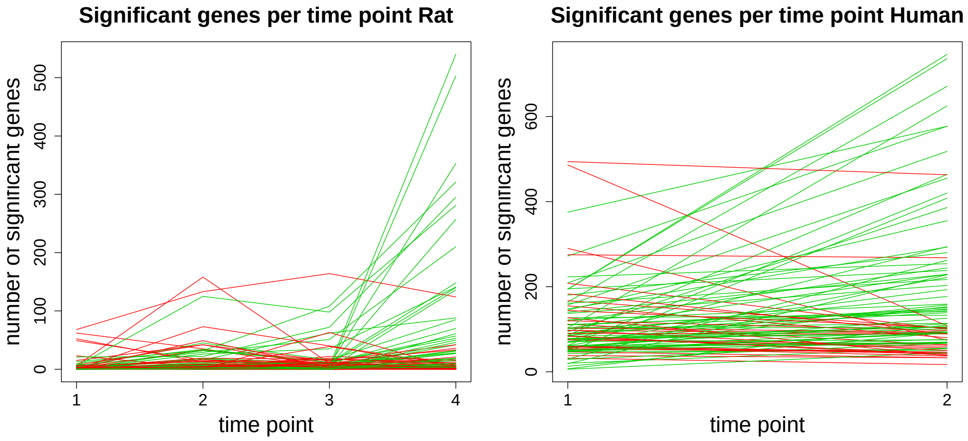


Figure 7.5: *Number of genes with significant dose-response relationship per time point. Green compounds have maximum in last time point, red compounds in any other time point. Left panel: Rat data results. Right panel: Human data results.*

represents differential expression of almost six-fold change increase, while isotonic mean for humans shows almost five log-fold change increase. Predictions of all dose effects in humans using high dose effect in rats, when only genes significant in rats are used, are explored in Figure 7.8. As expected, prediction of control dose shows very low correlation, since all values for human control dose should be around zero. However, for higher doses we can see that there are genes with (nearly) the same value of isotonic means both for rat and human. Still, there is large amount of genes centered around zero. However, in Figure 7.9, where only genes significant in both rat and human last time point are used, the subset of genes around zero almost disappears. The resulting gene set reveals genes that are both consistently significant across species and translatable between species with respect to fold change induced by high dose of a given compound (omeprazole in this case).

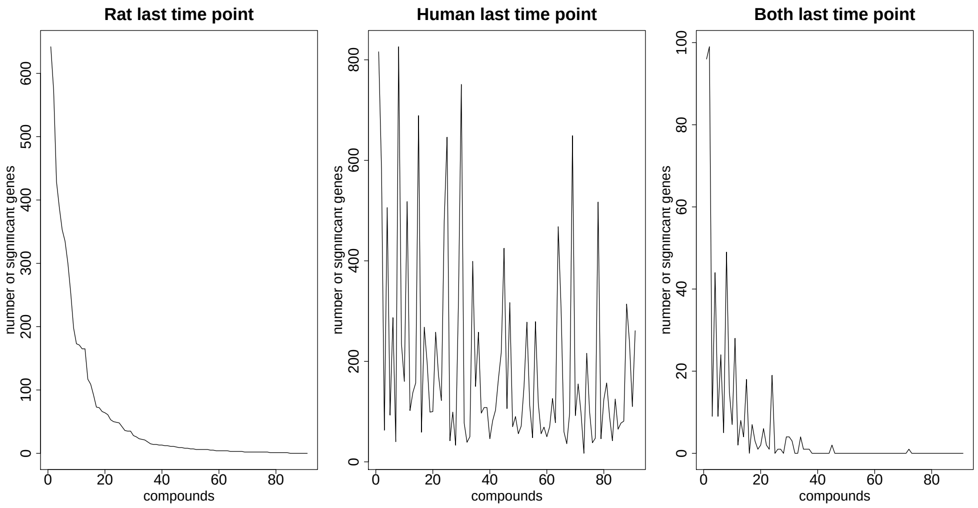


Figure 7.6: *Number of genes with significant dose-response relationship in last time point. Compounds are ordered according to the number of significant genes in rat and ordering is kept across all three panels. Rightmost panel is then intersection of two panels on the left.*

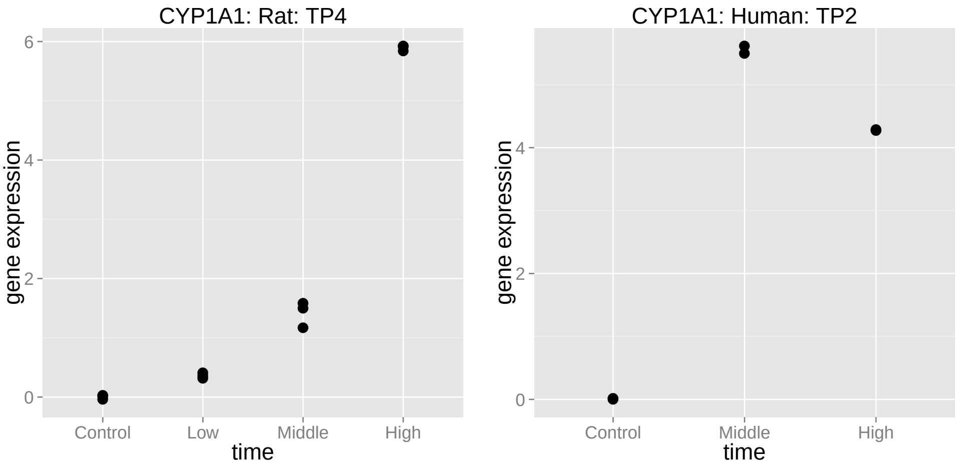


Figure 7.7: *Example of gene translatable between rat and human for compound omeprazole: gene* Cyp1a1 *in rat, respectively* CYP1A1 *in human.*

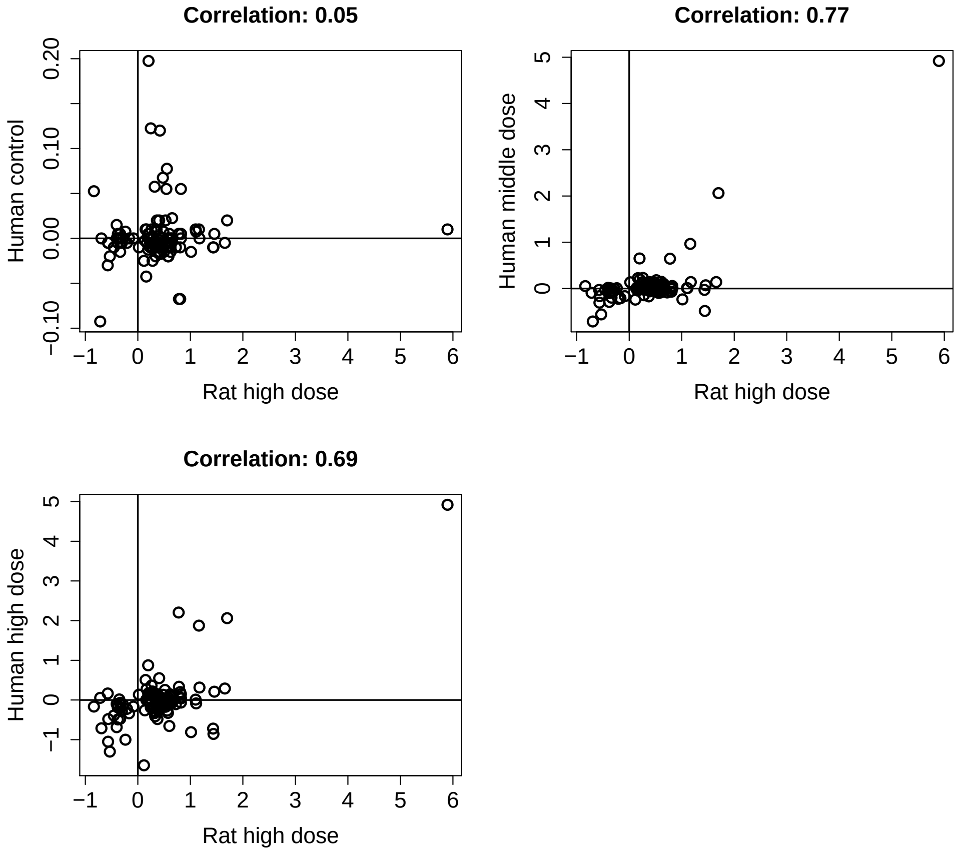


Figure 7.8: *Dose effect for the compound omeprazole: estimated isotonic mean in particular dose in human against estimated isotonic mean in high dose in rat, both for last time point. Genes with significant dose-response relationship for rat in last time point (significance in human is not considered).*

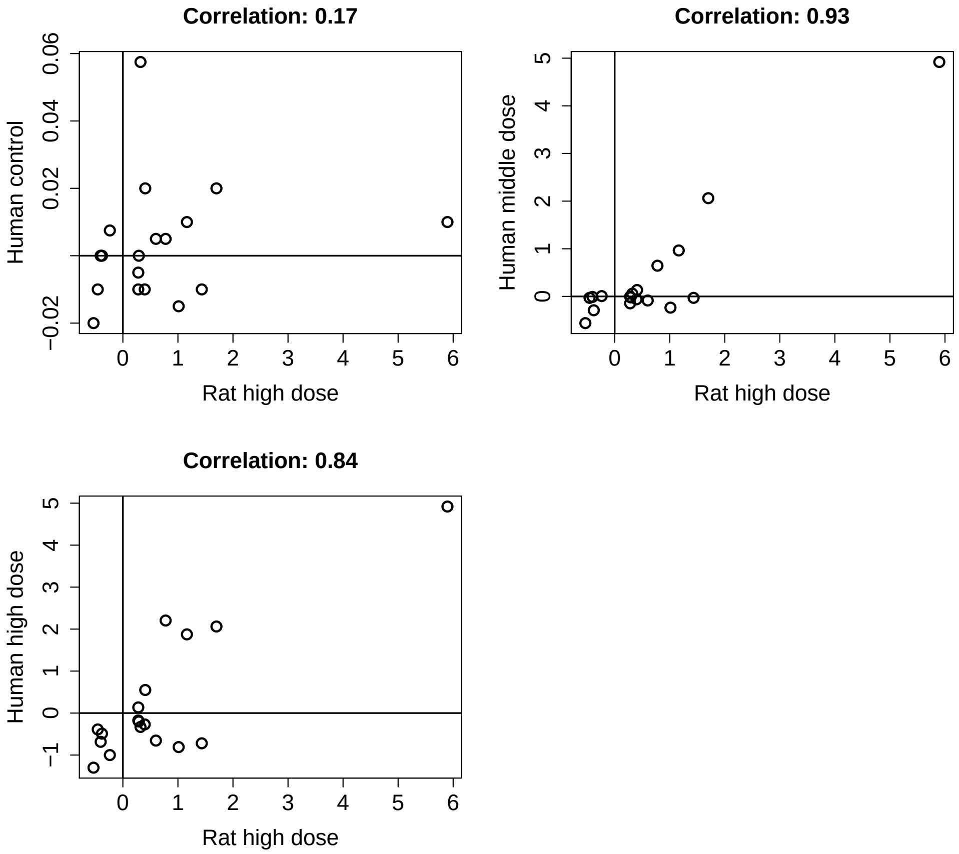


Figure 7.9: *Dose effect for the compound omeprazole: estimated isotonic mean in particular dose in human against estimated isotonic mean in high dose in rat, both for last time point. Genes with significant dose-response relationship for both rat and human in last time point.*

7.4 Discussion

According to the ANOVA results, the number of significant genes varied among the compounds. This finding is not surprising since the data set contains very distinct compounds, both with respect to their structural properties and biological effects. The data set contains vitamin A next to ibuprofen or nicotinic acid. The analyses presented in this chapter suggest that searching for overall differentially expressed genes can fail due to heterogeneity in the data set. Limiting ourselves to smaller subgroups of similar compounds can lead to more efficient analysis and meaningful results. One of such subsets was identified by our analysis, by grouping together 23 compounds with significant gene *ASF1A*. The presence of subgroups of compounds questions the meaningfulness of the goal of identifying genes useful for classification of compounds as DILI. If within given set of compounds would exist latent subgroups of compounds (similar with respect to their overall behavior), then particular genes could be good predictors of DILI in one subgroup, but not necessarily in the other subgroups. In other words, genes that can be predictors for DILI within one subgroup may lose its predictive ability by considering whole data set with several subgroups of compounds. Besides, the DILI response is highly unbalanced, only eight compounds out of 93 show "no DILI concern". Therefore, we propose to use a more specific response variable instead and simultaneously focus on possible identification of subgroups among compounds. These insights lead us to focus on translatability and means prediction in the second part of the analysis.

The second part of the analysis was mainly focused on the translatability of genes between humans and rats. The genes of interest are such that the fold change of their gene expression (precisely its log ratio against control) is similar in rat and human data and the dose-response relationship is statistically significant in both species. We have shown that for some compounds, no relevant results were found. This is mostly due to very low overall difference in expressions and high variability. However, for several compounds, we were able to identify such gene sets. The interpretation of the findings is clear: the value of gene expression observed in rats can be used as biomarker for the corresponding gene expression value in humans. If we are able to connect these genes with particular toxicological process, the signature made by these genes can serve as early warning mechanism. The reliability of such genes as biomarkers will need to be validated, but the fact that they are significant in both species may highlight a common underlying biological mechanism in both species after exposure to the compounds. This study may provide a leeway into more extensive studies on rats and humans toxicogenomics connectivity in early drug developments.

Chapter 8

Disconnected Genes in the Japanese Toxicogenomics Project

8.1 Introduction

The importance of translatability research is described in detail in Section 7.1.1. This chapter focuses on the translation from *in vitro* to *in vivo* within one species and it is relevant in both rat and human studies. We will explore frameworks to identify genes that shows discrepancies between rat *in vitro* and rat *in vivo* data. Identification of such genes could help to explain differences between processes in living animals and in cell cultures.

Zhang *et al.* (2014) developed consensus early response toxicity signatures of *in vitro* and *in vivo* toxicity in human and rat using time-dependent gene expressions. For the hepatotoxicant hydrazine, Timbrell *et al.* (1996) show that the effects on various parameters do not always show a quantitative or qualitative correlation between *in vivo* and *in vitro* data. Enayetallah *et al.* (2013) profiled nine compounds for *in vitro* and *in vivo* cardiotoxicity, and reported that while there were common biological pathways for *in vivo* and *in vitro* rat experiments for drugs like dexamethasone, most of the biological pathways identified *in vivo* for the drug amiodarone were not detected *in vitro*. Early prediction of safety issues for hit or lead compounds would benefit not only from consensus signatures, but also from *disconnect* signatures between *in vivo* and *in vitro* toxicogenomics experiments. These disconnect signatures can indicate which biological pathways are less likely to translate from a simplified *in vitro* model to a complex and holistic *in vivo* system.

Toxicity signatures developed from *in vitro* models most probably reflect protein modulations or pathway changes resulting from direct effects of compounds upon cells instead of the more complex interactions found in *in vivo* systems. *In vitro* signatures could also show excessive toxicity not to be detected *in vivo* due to compensatory mechanisms found in *in vivo* systems. Thus the framework is proposed to detect genes that are disconnected between *in vitro* and *in vivo* dose-dependent toxicogenomics experiments using fractional polynomial models. Biclustering is applied to find subsets of disconnected genes that are common to several compounds. Finally, the identified groups of disconnected genes are interpreted by their most probable biological pathways.

The Disconnect data described in Section 1.2.2.2 are used for the analysis in this chapter. Fractional polynomials and biclustering are introduced in Section 8.2. The analysis workflow is described and results are discussed in Section 8.3. Further integration of findings is explained in Section 8.4. The Section 8.5 summarizes the findings of the chapter.

8.2 Methods

A flexible fractional polynomial modelling framework is proposed to: (1) identify genes with significant dose-response relationships in an *in vitro* or *in vivo* experiments and (2) identify genes that are disconnected between the two systems. The *in vitro* and *in vivo* gene expression matrices were analysed jointly by compound and the resulting disconnected genes from the separate analyses were integrated using the *Bimax* biclustering algorithm (Prelic *et al.*, 2006) in order to identify subsets of disconnected genes that are common to several compounds.

8.2.1 The fractional polynomial framework

The fractional polynomial modelling framework aims to capture non-linear relationship between a predictor and a response variable. It assumes that most non-linear profiles can be captured by a combination of two polynomial powers (Royston and Altman, 1994). It is particularly appealing for modelling dose-response relationships since it does not impose monotonicity apparent in most dose-response modelling methods (e.g. Ramsay, 1988, Lin *et al.*, 2012d). For a single gene, let Y_{ij} denote gene expression *in vivo*, where $i = 1, 2, \ldots, m$ represents dose level and $j = 1, 2, \ldots, n_i$ denotes number of replicates per dose. The fractional polynomial framework assumes that the relationship between gene expression and doses can be captured by a polynomial function;

$$Y_{ij} = \beta_0 + \beta_1 \cdot f_{ij}(p_1) + \beta_2 \cdot g_{ij}(p_1, p_2) + \varepsilon_{ij}, \tag{8.1}$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$ and the polynomial powers $p_1, p_2 \in P$, $P = \{-3, -2.5, \ldots, 1.5, 2\}$, while $p_1 \leq p_2$. This range of values provides enough flexibility to capture different forms of dose-response profile (Royston and Altman, 1994). The functions $f_{ij}(p_1)$ and $g_{ij}(p_1, p_2)$ are defined as

$$f_{ij}(p_1) = \begin{cases} i^{p_1} & p_1 \neq 0, \\ \log(i) & p_1 = 0, \end{cases}$$

and

$$g_{ij}(p_1, p_2) = \begin{cases} i^{p_2} & p_2 \neq p_1,\ p_2 \neq 0, \\ \log(i) \cdot i^{p_2} & p_1 = p_2,\ p_2 \neq 0, \\ \log(i) & p_2 \neq p_1,\ p_2 = 0, \\ \log(i) \cdot \log(i) & p_2 = p_1 = 0. \end{cases} \tag{8.2}$$

Note that for $p_1 \neq 0$, $p_2 \neq 0$ and $p_1 \neq p2$, the fractional polynomial model is given by $Y_{ij} = \beta_0 + \beta_1 \cdot i^{p_1} + \beta_2 \cdot i^{p_2} + \varepsilon_{ij}$. An example of fitting different combinations of powers for one particular gene is shown in Figure 8.1.

Akaike's information criterion (AIC, Akaike, 1974) is used to select the optimal combination of p_1 and p_2 that best reflects the observed dose-response relationship. Optimal solutions are denoted by $\{\hat{\phi}_1, \hat{\phi}_2\} = \Big\{\{p_1, p_2\} \in P, \mathrm{AIC}(\hat{\phi}_1, \hat{\phi}_2) = \min[\mathrm{AIC}(p_1, p_2)]\Big\}$. In order to identify genes with a significant dose-response relationship *in vitro*, a likelihood-ratio test (LRT, Neyman and Pearson, 1933) is used to compare model (8.1) that best fits the data and model (8.3), the null model that assumes no dose effect:

$$Y_{ij} = \beta_0 + \varepsilon_{ij}. \tag{8.3}$$

This additional testing is necessary in order to identify genes with statistically significant difference from the null model.

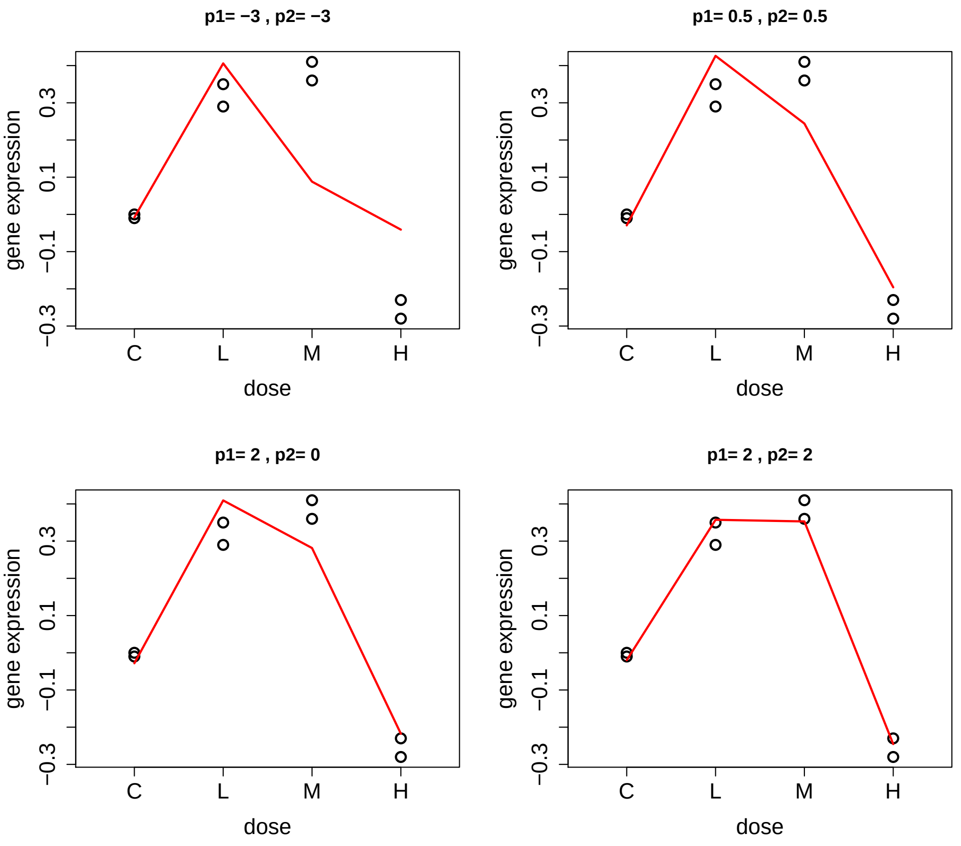


Figure 8.1: *Gene A2m for compound sulindac. Different combinations of powers are used and the model is fitted to the data (red solid line). The model in top left panel does not follow the data very well, the model in the bottom right panel is the best fitting model, given* $p_1, p_2 \in \{-3, -2.5, \ldots, 1.5, 2\}$.

To identify disconnected genes when comparing *in vitro* and *in vivo* data, the optimal fractional polynomial function selected per gene (with $\hat{\phi}_1, \hat{\phi}_2$, as fixed above) from *in vitro* data set is projected to *in vivo* data set under the assumptions that both *in vitro* and *in vivo* dose-response relationships are similar. For a single gene, let X_{ijk} denote gene expression *in vitro* and *in vivo*, where $i = 1, 2, \ldots, m$ represents dose levels, $j = 1, 2, \ldots, n_i$ denotes number of replicates per dose and $k = 1$ or $k = 2$ depending on whether the data is from *in vitro* or *in vivo* experiment. The *in vitro* - *in vivo* projected fractional polynomial model is specified as

$$X_{ijk} = \beta_0 + \beta_1 \cdot f_{ijk}(\hat{\phi}_1) + \beta_2 \cdot g_{ijk}(\hat{\phi}_1, \hat{\phi}_2) + \varepsilon_{ij}, \tag{8.4}$$

where $\varepsilon_{ijk} \sim N(0, \sigma^2)$. A LRT is used to quantify the dissimilarity in *in vivo* - *in vitro* dose-response relationships. It compares model (8.4), which assumes that dose-response relationships from *in vitro* and *in vivo* experiments are the same, with model (8.5), which assumes different dose-response relationships.

$$X_{ijk} = \begin{cases} \beta_0 + \beta_1 \cdot f_{ijk}(\hat{\phi}_1) + \beta_2 \cdot g_{ijk}(\hat{\phi}_1, \hat{\phi}_2) + \varepsilon_{ijk} & \textit{in vitro}, \\ \\ (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1) \cdot f_{ijk}(\hat{\phi}_1) + (\beta_2 + \gamma_2) \cdot g_{ijk}(\hat{\phi}_1, \hat{\phi}_2) + \varepsilon_{ijk} & \textit{in vivo}. \end{cases} \tag{8.5}$$

The comparison translates into testing if $\gamma_0 = \gamma_1 = \gamma_2 = 0$ in model (8.5). An example of a projected fractional polynomial model is shown in Figure 8.2. A significant result obtained from LRT comparison of model (8.4) and model (8.5) can be interpreted as a disconnect in gene expression between *in vitro* and *in vivo* rat experiments. The significance level was specified as 10% after correction for multiplicity (Benjamini and Hochberg, 1995). Resulting disconnected genes were subjected to fold change filtering by excluding genes with maximal dose-specific fold change between *in vitro* and *in vivo* data set less than 1. The fold change filtering further reduces false positives due to small variance genes (Talloen and Göhlmann, 2009).

The empirical validation of the method in the context of *in vitro* and *in vivo* disconnects was done through a series of simulation studies. In summary, the proposed projected fractional polynomial method under the null model resulted in 90% specificity using the same number of dose and the same number of observations per dose as in TGP data set. When number of observations per dose was increased to four, specificity increased up to 98%. Under the alternative hypothesis of a disconnected dose-response profiles between *in vitro* and *in vivo* experiments, the method resulted in 100% sensitivity for the disconnected linear profiles. For nonlinear profiles, sensitivity of 80% - 95% was achieved, for the maximum fold change between the *in vitro* and *in vivo* settings greater than 1.2. Sensitivity increased up to 98% - 100% when the fold change was greater than 1.6. The method also resulted in 93% specificity and 95% sensitivity after multiplicity correction. The simulation studies indicated that the method may perform better in other settings than the reported results for the TGP experiment due to its limited number of replicates per dose and the weak signals. The full description of the simulations' settings and results can be found in the Appendix.

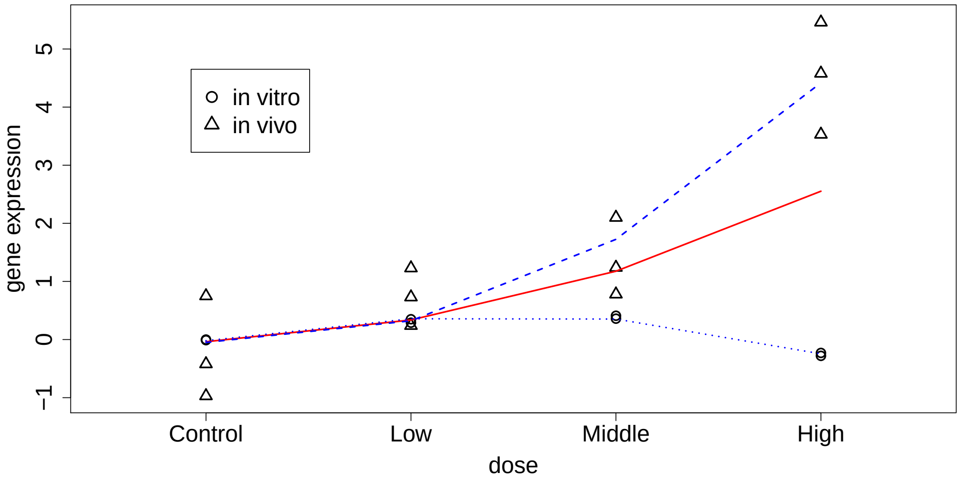


Figure 8.2: *Gene A2m for compound sulindac. Consequences of forcing the same model to both data sets. Red solid line shows the profile, if both data sets share parameters, i.e. model (8.4), and blue lines show fits for model (8.5), i.e. we consider the same powers but separate parameters for* in vitro *(dotted line) and* in vivo *data (dashed line). Circles represent* in vitro *and triangles* in vivo *data. Clearly, for this particular gene, model (8.5) provides better fit.*

8.2.2 Biclustering of genes and compounds

A biclustering framework was introduced in order to find subsets of genes and conditions with a similar pattern (Cheng and Church, 2000). Biclustering methods (Madeira and Oliviera, 2004, Eren *et al.*, 2013) are designed to cluster in two dimensions simultaneously to produce sub-matrices of the original data that behave consistently in both dimensions. The resulting sub-matrices are called biclusters. Based on the identified disconnected genes from the fractional polynomial models, a disconnect matrix $D_{(G\times C)}$ of binary values was created with element d_{gc} defined as:

$$d_{gc} = \begin{cases} 1 & \text{if gene } g \text{ is disconnected for compound } c, \\ 0 & \text{otherwise,} \end{cases} \tag{8.6}$$

where G is the number of genes that are significant for at least one compound (i.e. $G \leq 5,914$) and $C = 128$ is the number of compounds. The Bimax algorithm (Prelic *et al.*, 2006) for binary data is applied to the disconnect matrix (G) to find subsets of the disconnected genes that are common to several compounds.

8.3 Results

The data were analysed in two ways depending on the direction of the projected fractional polynomial models. The first set of models (*in vitro disconnects*) defined the fractional polynomial powers based on the *in vitro* data set and projected its dose-response profiles to the *in vivo* data set. The second set of models (*in vivo disconnects*) defined the fractional polynomial powers based on the *in vivo* data set and projected its dose-response profiles to the *in vitro* data set. The analyses were performed in statistical software `R` version `3.0.1` (R Core Team, 2013).

8.3.1 *in vitro* disconnects

The final set of disconnect genes, identified using the fractional polynomial model, consists of 3,348 genes that were disconnected for at least one compound. The number of the identified disconnected genes per compound ranged from zero to 1,276 (with median 37.5), with maximal value for compound colchicine. There were ten compounds with no disconnected genes and an additional 27 compounds with less than ten genes. There are 1,022 genes that are disconnected only for a single compound. Three genes (*Aldh1a1*, *Cyp1a1* and *Fam25a*) were consistently identified in 56 compounds while 446 genes were detected in more than ten compounds. The 446 genes were analysed further for common biological pathways using GO (Ashburner *et al.*, 2000), KEGG (Kanehisa

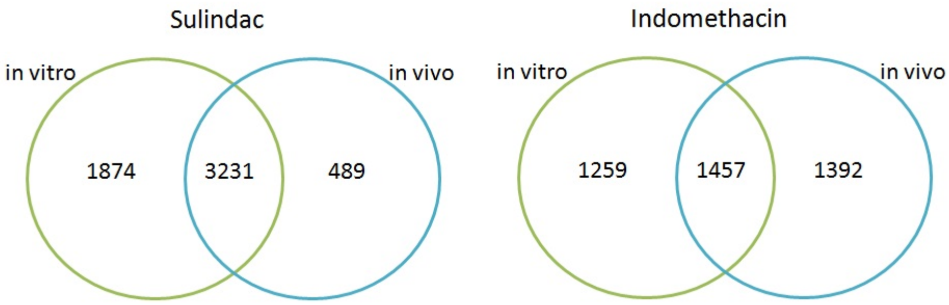


Figure 8.3: *Number of genes with significant dose-response relationship for* in vitro *data only, for* in vivo *data only and for both data sets simultaneously. The sum of all the numbers gives us the number of genes significant at least for one data set in a given compound. Two compounds are shown as examples.*

and Goto, 2000) and *Janssen pharmaceutica* in-house gene databases. As expected, many of the genes are involved in drug metabolism (e.g. acetaminophen metabolism, Benzo[a]pyrene metabolism, CAR/RXR activation, PXR/RXR activation), as well as endogenous compound metabolism (e.g. butanoate metabolism, alanine, cysteine and methionine metabolism, nitrogen metabolism, fatty acid metabolism, cholesterol biosynthesis). Additionally, some of the genes are also involved in toxicity related pathways such as oxidative stress due to reactive metabolites, bilirubin increase, glutathion depletion and phospholipidosis as well as complex pathways such as immune response, classical complement and coagulation. Only pathways containing more than five genes and with coverage of more than 10% (i.e. more than 10% of their genes were disconnected genes) were considered.

The biclustering Bimax procedure was applied on binary matrix $D_{(3348\times 128)}$, with a minimal bicluster size of four compounds. We identified 188 unique genes that were consistently defined as disconnected genes in seven compounds based on the first ten biclusters from the Bimax algorithm (left panel on Figure 8.4). Sulindac and diclofenac are both anti-inflammatory drugs, acetic acid derivatives that are likely to damage liver (Rodríguez *et al.*, 1994). Naphthyl isothiocyanate was shown to cause direct hepatotoxicity (Williams, 1974). Among the 188 genes, the top genes (with respect to fold change) are associated with liver toxicity. Genes *A2m* and *Lcn2* were validated for being affected in case of hepatotoxicity (Wang *et al.*, 2008). Other toxicity associated genes found were *Cyp1a1*, *Pcsk9*, *Car3*, *Gstm3* or *Ccnd1*. Table 8.1 shows the results of pathway analysis for the first bicluster (compounds: sulindac, naphthyl isothiocyanate, diclofenac and

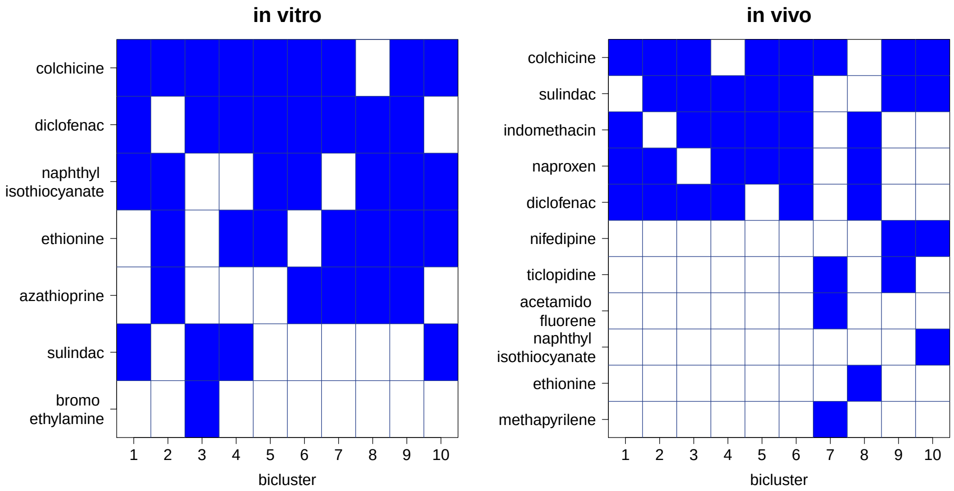


Figure 8.4: *Appearance of compounds across ten biclusters. Blue colour states that the compound is the member of bicluster. Left panel: Analysis starting with* in vitro *data. Right panel: Analysis starting with* in vivo *data.*

Table 8.1: The genes showing disconnect that are members of bicluster 1 and their membership in pathways. The pathways were identified using KEGG (Kanehisa and Goto, 2000).

Pathway	Genes
Complement and coagulation cascades	*A2m C1s C5 C8a C4bpb Cfh F*5
Chemical carcinogenesis	*Cyp1a1 Gstm3 Gsta5*
Metabolism of xenobiotics	*Akr7a3 Cyp1a1 Gstm3 Gsta5*
Pathways in cancer	*Ccnd1 Fn1 Lamc2*

colchicine). Genes *A2m*, *Gpx2* and *Gstm3* were disconnected genes common to all the seven compounds and other 16 genes (e.g. *C5*, *Fam25a*, *Gsta5*) appeared for six of them simultaneously.

8.3.2 *in vivo* disconnects

The final set of disconnect genes contained 2,346 genes that were disconnected *in vivo* for at least one compound. The number of the identified disconnected genes per compound ranged from zero to 798 (with median 18), with maximal value for compound colchicine. There were 25 compounds with no disconnect gene and another 29 with less than ten genes. There were 992 genes that appeared only for a single compound. The gene *Stac3* showed disconnect for 54 compounds.

There were 175 genes that showed disconnect in gene expression from *in vivo* to *in vitro* rat experiments for more than ten compounds. Similar pathways as in the previous section (i.e projection from *in vitro* to *in vivo*) were also discovered, although more of the pathways were related to exogenous compound metabolism. Oxidative stress endpoints related pathways were more common *in vivo*. Complex pathways such as complement and coagulation identified in the *in vitro* data set were not discovered in the analysis of the *in vivo* data set, which may be due to differences between the prescribed dose and actual exposure in liver tissue *in vivo*.

The Bimax algorithm as applied on the binary matrix $D_{(2346\times128)}$, with a minimal cluster size of four compounds. It identified 163 unique genes common to 11 distinct compounds based on the first ten biclusters (right panel on Figure 8.4). Five compounds were identified both in *in vitro* and *in vivo* analyses of disconnects: sulindac, colchicine, diclofenac, ethionine and naphthyl isothiocyanate. The most interesting of the additional compounds are indomethacin and naproxen. They are both members of a group of non-steroidal anti-inflammatory drugs (NSAIDs), the former an acetic acid derivative and the latter a propionic acid derivative. Both drugs are nonselective cyclooxygenase (COX) isozyme inhibitors, i.e. with undesired targeting of COX-1 that leads to gastrointestinal adverse effects (Rao and Knaus, 2008; Brune and Patrignani, 2015). Specifically, both drugs are indicated as high risk drugs for general upper gastrointestinal complications (Castellsague *et al.*, 2012). All of the compounds are connected to toxicity events. Most of the toxicity related genes (*A2m*, *Lcn2*, *Car3*, *Pcsk9*, *Acsl1*, *Lamc2*, *Selenbp1* and *Serpina10*) from the previous *in vitro* analysis were also identified from the analysis of the *in vivo* data set. Other toxicity related genes were *Cyp2e1* (Heijne *et al.*, 2005), *Upp1*, *Gss*, *Ddc*, *Gstm7* and *Srebf1*. One gene was disconnected between *in vitro* and *in vivo* for all the 11 compounds (*A2m*) and additional four genes appeared for more than eight compounds simultaneously (*Scd1*, *Srebf1*, *Stac3*, *Xpnpep2*).

8.4 Discussion

The analytical framework identified three broad groups of genes from a joint analyses of *in vitro* and *in vivo* rats toxicogenomic experiments. The first group showed a significant dose-response relationship in both the *in vitro* and *in vivo* toxicogenomic experiments (478 genes for sulindac, e.g. *A2m*, *Car3*, *Lcn2*). These types of genes are shown in the top panels of Figure 8.5. While some of the genes in this group showed contradictory dose-responses profiles between the *in vitro* and *in vivo* data, others showed similar dose-response profiles, but with different magnitude of gene expression values. The second group contains genes that showed a significant dose-response relationship in *in vitro* experiments, but not in *in vivo* experiments (205 genes for sulindac, e.g. *Cd44*, *Gstm3*, *Gsta5*). Examples of such genes are presented in the top panels of Figure 8.6. This set of genes may represent the difference in biological complexity between *in vivo* and *in vitro* systems. The third group are those genes that showed significant dose-response relationship in *in vivo* experiments, but not in *in vitro* experiments (30 genes for sulindac, e.g. *Akr1c3*, *Cyp2a2*, *Scd1*). This set of genes may occur due to the mechanism of action (MoA) *in vitro* of a drug candidate not being representative of *in vivo*. Examples of such genes are presented in the bottom panels of Figure 8.6.

Most of the compounds in our specific case study that triggered the expression of the identified disconnected genes are members of a group of anti-inflammatory drugs and all of them are related to hepatotoxicity, nephrotoxicity or gastro-intestinal toxicity. Genes that were shared across compounds were related to toxicity, drug metabolism and liver or kidney development. In total, there were 188 genes discovered by the *in vitro* analysis (e.g. *Gsta5*, *Gstm3*) and 163 genes by the *in vivo* analysis (e.g. *Ddc*, *Scd1*), focusing on first 10 biclusters. Highly relevant may be the 63 genes (e.g. *A2m*, *F5*, *Lcn2*) that were found by both analyses, i.e. showing disconnect while having significant dose-response relationships both *in vitro* and *in vivo*.

If additional data about experiments are available both for *in vitro* and *in vivo*, such data can be included in the proposed methodology. The adjustment can be done by adding the new variables in the fractional polynomial model as covariates. Note that in this type of gene expression studies, the rats are specially bred to ensure baseline comparability across all rats.

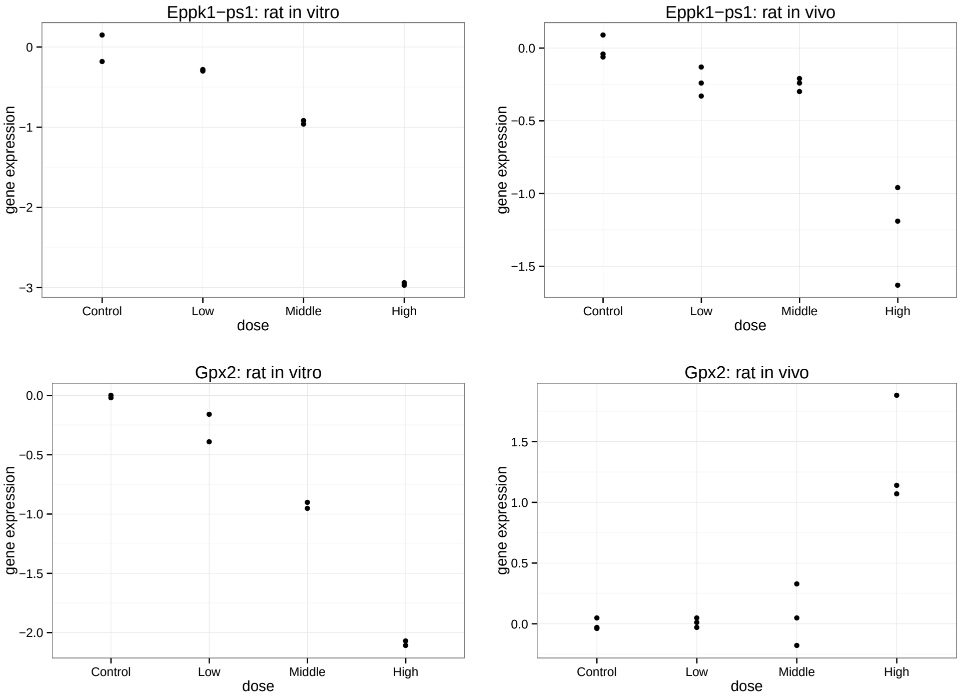


Figure 8.5: *Example: compound sulindac. Two genes from Group 1. Top panels: gene Eppk1-ps1 with same direction, but different magnitude of effect. Bottom panels: gene Gpx2 with different direction of effect across systems. Left panels:* in vitro. *Right panels:* in vivo.

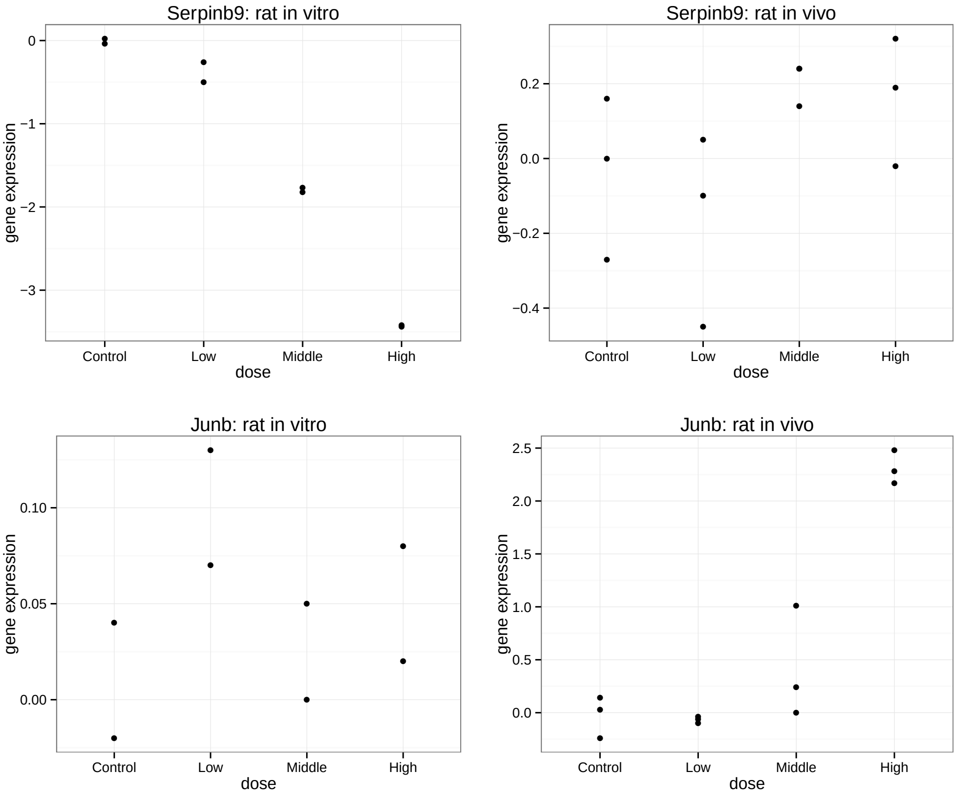


Figure 8.6: *Example: compound sulindac. Top panels: gene Serpinb9 from Group 2, with effect only in* in vitro*. Bottom panels: gene Junb from Group 3, with effect only in* in vivo*. Left panels:* in vitro*. Right panels:* in vivo.

8.5 Conclusion

The findings demonstrated that substantial differences may be identified between *in vitro* and *in vivo* gene expression. While this result is not surprising, the importance of the analysis is in the identification of different groups of the disconnected genes. Genes were identified that showed significant dose-response relationships with compounds *in vitro* that were absent *in vivo*, and vice-versa. Moreover, there was a group of genes with a different direction of dose-response relationship across the two systems. These finding confirms possibility of important discrepancies between *in vitro* experiments and *in vivo* experiments. Pathway analysis of the identifying disconnected genes between *in vivo* and *in vitro* rat system may improve our understanding of uncertainties in mechanism of action of a drug candidate in human, especially for early toxicity detection.

Part III

Software Development for Dose-response Omics Data

Chapter 9

Order Restricted Clustering for Microarray Experiments

9.1 Introduction

Dose-response analysis of microarray data is a fast growing area of scientific interest. According to Ernst and Bar-Joseph (2006), in 39.1% of the 786 data sets in the Gene Expression Omnibus of 2005 are studies with an ordered restricted design variable such as age, time, temperature and dose. Among these data sets, 1% are dose-response studies. Table 9.1 presents a list of free software developed for the analysis of gene expression experiments with an order restricted design.

There is a substantial amount of overlap between the different packages presented in Table 9.1 and the same or a similar analysis can be conducted using more than one package. In R (R Core Team, 2013), there are several packages available. IsoGene (Lin *et al.*, 2013) and IsoGeneGUI (Pramana *et al.*, 2012a; more detail about both packages in Pramana *et al.*, 2010 and Lin *et al.*, 2012b) are CRAN R packages which can be used for inference and data exploration of dose-response microarray data. ORIClust (Liu *et al.*, 2012) is a CRAN package for clustering of time-series and dose-response microarray data (Liu *et al.*, 2009 and Lin *et al.*, 2009) using order restricted information criteria. Package orQA (Klinglmueller, 2010) is a CRAN package for inference of order restricted for cross platform microarray data. The ORIOGEN package (Peddada *et al.*, 2003) is a Java-based (Arnodl *et al.*, 2000) interface which can be used for both inference and clustering of dose-response and time-series data. In this chapter we present new methodology for two stage clustering of dose-response microarray data under order restriction. This novel methodology is based on δ-clustering and has been implemented in the CRAN package ORCME (Otava *et al.*, 2014). The method and package are applicable in general framework of order restriction, but main focus (and specific package functions) are related to the special case, when monotone profiles are of primary interest.

In Section 9.2, we briefly review the original δ-biclustering method (Cheng and Church, 2000) and derived δ-clustering method for whole profiles clustering. The ORCME R package is introduced in Section 9.3 and the use of the package is illustrated for a case study of dose-response microarray data. The options for choice of homogeneity parameter are described in Section 9.4 and the chapter is summarized in Section 9.5.

Package	Type	Location	Reference
orQA	R	CRAN	Klinglmueller *et al.* (2011)
IsoGene	R	CRAN	Pramana *et al.* (2010)
IsoGeneGUI	R	Bioconductor	Pramana *et al.* (2012a)
ORIOGEN	Java	website	Peddada *et al.* (2003)
ORIClust	R	CRAN	Liu *et al.* (2009)
STEM	Java	website	Ernst and Bar-Joseph (2006)
ORCME	R	CRAN	Otava *et al.* (2014)

Table 9.1: *Software for dose-response and time course gene expression data.*

9.2 Order restricted curve clustering

Denote a gene expression matrix $\boldsymbol{Y}$, with dimension $M \times I$, where number of genes and conditions are denoted with M and I, respectively. The matrix entries are denoted as y_{mi}, where the index represents mth gene under condition (dose) i. Note that there is only one entry per gene and dose combination. In case of replicates for the dose level, the y_{mi} represents the mean value. Define y_{MI} as the overall mean of the expression matrix $\boldsymbol{Y}$, y_{mI} as the mean expression of gene m and y_{Mi} is the mean expression of condition i. In general, we assume some order restriction assumption about the y_{mi} in the sense of the increasing dose i. Specifically, we further assume the monotonicity of dose-response relationship.

The two stage δ-clustering procedure discussed in following sections consist of implementing the order restrictions and clustering itself. Prior to the method itself, the inference-based filtering should be applied. The initial filtering step is necessary in order to discard non-significant genes. The within gene variability is ignored by the δ-clustering method and the clusters are constructed in order to reduce the between gene variability (i.e. the within cluster variability). Without filtering, the non-significant genes (with high within-gene variability) would enter the clusters and interpretation would be compromised. The filtering step could be done with any suitable method like t-test type statistics: William's (Williams, 1971 and Williams, 1972), Marcus' (Marcus, 1976), Hu's (Hu *et al.*, 2005) and modified Hu's (Lin *et al.*, 2007) test statistics or likelihood-ratio test discussed by Bartholomew (1961), Barlow *et al.* (1972), and Robertson *et al.* (1988). In the examples discussed below, we use the likelihood-ratio test that compares the ratio between the variance calculated under the null hypothesis (the constant dose-response profile) and the variance calculated under an ordered alternative. In case of a significant test result, we can straightforwardly derive the direction of the monotonicity by comparing the likelihood under upward or downward monotone alternative.

9.2.1 The δ-biclustering method

The δ-biclustering is a node deletion based algorithm introduced by Cheng and Church (2000) to find a subset of genes and conditions with a high similarity score. The similarity between members of a bicluster is defined in terms of the mean squared residue score. The lower the mean squared residue score, the more homogeneous is the cluster. The δ-biclustering method relies on the assumption that every entry in a gene expression matrix can be expressed in terms of its row mean, column mean, the overall mean of the expression matrix and random error. Hence, the residue of expression value of the mth gene under condition (dose) i can be expressed as:

$$r_{mi} = y_{mi} - y_{Mi} - y_{mI} + y_{MI}, \tag{9.1}$$

and the mean squared residue score of matrix $\boldsymbol{Y}$ and of gene m is defined as:

$$H(\boldsymbol{Y}) = \frac{1}{MI}\sum_{m=1}^{M}\sum_{i=1}^{I} r_{mi}^2 \qquad d_m(\boldsymbol{Y}) = \frac{1}{I}\sum_{i=1}^{I} r_{mi}^2.$$

Note that the model for the residual in Equation (9.1) can be expressed in the form of a two-way ANOVA model without an interaction term:

$$y_{mi} = \mu + \alpha_m + \beta_i + r_{mi}, \tag{9.2}$$

with $\mu = y_{MI}$, $\alpha_m = y_{mI} - y_{MI}$ and $\beta_i = y_{iM} - y_{MI}$.

As an illustration, we present an example of two expression matrices. Matrix $\boldsymbol{A}$ is an example of a perfect cluster with coherent values and $\boldsymbol{B}$ is an example of a cluster for which the genes have coherent values except for the genes in the last two rows of the matrix. Based on Equation (9.1), the mean squared residue score for $\boldsymbol{A}$ is zero since the total variability of the cluster can be explained by the row means, column means and overall mean of the matrix. However, for $\boldsymbol{B}$ the mean squared residue score is 8.11. This means that genes in $\boldsymbol{A}$ are more similar than those in $\boldsymbol{B}$. Suppose that the last two rows of $\boldsymbol{B}$ are excluded, then the mean squared score becomes zero.

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 6 \\ 30 & 31 & 32 & 33 & 34 \\ 32 & 33 & 34 & 35 & 36 \\ 81 & 82 & 83 & 84 & 85 \\ 91 & 92 & 93 & 94 & 95 \end{pmatrix}$$

$$\boldsymbol{B} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 6 \\ 30 & 31 & 32 & 33 & 34 \\ 32 & 33 & 34 & 35 & 36 \\ 42 & 43 & 30 & 30 & 31 \\ 37 & 30 & 36 & 35 & 34 \end{pmatrix}$$

In microarray experiments, a perfect cluster/bicluster such as $\boldsymbol{A}$ is unlikely given the noise level of the technology. It may therefore be sufficient to find clusters/biclusters of genes whose mean squared residue scores are less than a pre-specified threshold δ. Cheng and Church (2000) proposed the δ-biclustering method for gene expression data based on a suit of node deletion algorithms that evolve in cycles. The algorithm starts from the input gene expression matrix until a bicluster that satisfies the δ-criterion is found. Then the members of this cluster are replaced with the random data and the node deletion is applied again until another bicluster satisfying δ-criterion is found. Several cycles of the algorithm are then applied to the data by replacing the found biclusters with random data at the end of every cycle.

9.2.2 The δ-clustering of order restricted dose-response profiles

The goal of biclustering is to find subset of genes behaving similarly on subset of conditions. However, in the usual experimental settings, the column effects β in Equation (9.2) have an inherent ordering, which may be due to time, temperature, or, as in our example, increasing doses of a therapeutic compound. The aim is to find clusters of genes that have similar profiles represented by their dose-specific means. Therefore, the clustering of whole profiles is of interest, rather than clustering according to subset of condition (which is output of biclustering procedures). To achieve this goal, we propose the δ-clustering, a variant of δ-biclustering of Cheng and Church (2000).

First, we will explain the methodology of δ-clustering and then we will relate it to the microarray experiment data. Note one important difference between the aim of biclustering and clustering methods. If subsets are of interest (biclustering), appearance of genes in more clusters is necessary property. Gene can show similarity to varying groups of genes, depending on subset of interest. However, if whole profiles are of interest (clustering), gene should be member of one cluster of similar genes. Therefore, we are searching for non-overlapping clusters. One consequence for procedure would be rather deleting of already clustered genes from gene expression matrix then replacing them with random numbers, as is done in original δ-biclustering method.

9.2.2.1 The δ-clustering method

Applying the δ-biclustering algorithm in only one dimension offers a δ-clustering method for which the number of clusters is not required to be specified but implicitly controlled by the degree of homogeneity assumed for a cluster. However, the choice of a δ value to achieve a desired degree of homogeneity is not readily available (Prelic *et al.*, 2006). We propose a relative δ criterion, where a cluster is a subset of genes with a mean squared residue score smaller than a certain proportion λ $(0 \leq \lambda \leq 1)$ of the heterogeneity in the observed data. Searching for cluster consists of two steps. First, single node deletion algorithm is applied until heterogenous cluster is found, then nodes addition step is performed to form final cluster (Cheng and Church, 2000). Deletion is based on gene specific mean squared residue $d_m(\boldsymbol{Y})$, the gene with highest d_m is deleted in each step. Node deletion is stopped, when the mean squared residue H of remaining genes is smaller than $\lambda H(\boldsymbol{Y})$. Due to nature of algorithm, some genes that actually fit in the resulting cluster could have been thrown away during node deletion procedure. Therefore, we add back the genes for which d_m computed under reduced matrix $\boldsymbol{Y}^*$ is smaller than $\lambda H(\boldsymbol{Y}^*)$. Then, the enriched cluster is considered complete. Before proceeding to search for another cluster, the genes already clustered are omitted from the gene expression matrix. This is the consequence of our interest in non-overlapping clusters. The procedure is described in Algorithm 1 and mathematical details of the algorithm can be found in Cheng and Church (2000).

To overcome the problem of local minima (Prelic *et al.*, 2006), we introduce an additional parameter ϕ that indicates the minimum number of genes in a cluster. Note that for $\lambda = 0$, the algorithm searches for clusters of genes with mean squared residue score of zero, which may result in as many clusters as the number of genes in the data set. On the other hand, specifying λ to be one means to consider all the genes as one cluster. Any value of λ between zero and one reflects the degree of homogeneity expected of a cluster. We define the algorithm to carry out this task as Algorithm 1.

Algorithm 1: δ-clustering

Input: $\boldsymbol{Y}$, a matrix of real numbers; ϕ, minimum number of genes in a cluster; and λ: $0 \leq \lambda \leq 1$.

Output: Set of K clusters $\boldsymbol{Y}_k^{sub}$, $k = 1, \ldots, K$. Clusters are sub-matrices with number of rows smaller than or equal to the number of the rows of the original matrix $\boldsymbol{Y}$. The number of columns stays fixed, because we focus on whole profiles clustering.

Initialization: $\delta = \lambda \cdot H(\boldsymbol{Y})$, where $H(\boldsymbol{Y})$ is the mean squared residue score of the observed data. Set $\boldsymbol{Y}_A = \boldsymbol{Y}$ and $k = 1$.

Iteration:

1. Define $\boldsymbol{Y}_k^{sub} = \boldsymbol{Y}_A$.
2. Denote dimension of $H(\boldsymbol{Y}_k^{sub})$ as $P \times I$, where $P \leq M$. If $H(\boldsymbol{Y}_k^{sub}) < \delta$ or $P \leq \phi$, output $\boldsymbol{Y}_k^{sub}$ to step 4.
3. Perform single node deletion step: delete gene with highest d_m. Go to step 1 with new (reduced) $\boldsymbol{Y}_k^{sub}$ of row dimension $P - 1$.
4. Perform node addition step: add genes to $\boldsymbol{Y}_k^{sub}$ if for their d_m computed under $H(\boldsymbol{Y}_k^{sub})$ holds that $d_m \leq H(\boldsymbol{Y}_k^{sub})$. Output updated $\boldsymbol{Y}_k^{sub}$ as found cluster.
5. Update matrix $\boldsymbol{Y}_A$ by deleting all the genes that are members of cluster $\boldsymbol{Y}_k^{sub}$. If the matrix $\boldsymbol{Y}_A$ is not empty, set $k = k + 1$ and go back to step 1 with new matrix $\boldsymbol{Y}_A$.

Note that the Algorithm 1 allows to cluster subsets of genes with similar dose-response curve shapes. It is fairly general and it can be applied to any setting of an ordered design variable (time, temperature, dose etc.). It does not require particular order restrictions, such as monotone gene expression profile. The order restriction has to be built in within the first stage of our two-stage algorithm. In the following section, we discuss an algorithm that will incorporate the monotonicity assumption through isotonic regression (Robertson *et al.*, 1988). Consequently, we would be able to cluster together genes with similar monotone dose-response curve shapes. Note that the δ-clustering algorithm is usually applied to an expression matrix after an initial filtering where genes with no significant dose-response relationship are excluded from the analysis.

9.2.2.2 The δ-clustering of dose-response monotone profiles

A typical dose-response microarray data $\boldsymbol{Y}$ has entries y_{mij} corresponding to the expression level of gene m under dose i from subject/sample j. Usually, different subjects/samples are used for different doses, denoted N_{mi}. Since only single value per gene-dose level combination is considered for Algorithm 1, the dose-specific means are computed for each particular gene as

$$y_{mi} = \sum_{j=1}^{N_{mi}} \frac{y_{mij}}{N_{mi}}.$$

Computation of means is the moment when order restrictions are incorporated in the procedure. In order to find clusters of genes with a similar monotone dose-response relationship, it is required that gene expression means under increasing doses are constrained to be monotone. The isotonic regression (Robertson *et al.*, 1988) is used if monotone means are of interest. A new matrix $\boldsymbol{Y}^*$ of the isotonic means is obtained. The effect of the mth row (gene) α_m, the isotonic effect of the ith column (dose) (β_i^*) and the overall mean (μ) can be defined as shown below:

$$\mu = \textstyle\sum_{m=1}^{M} \sum_{i=1}^{I} \frac{y_{mi}^*}{MI},$$

$$\alpha_m = \textstyle\sum_{i=1}^{I} \frac{y_{mi}^*}{I} - \mu,$$

$$\beta_i^* = \textstyle\sum_{m=1}^{M} \frac{y_{mi}^*}{M} - \mu.$$

The clustering algorithm is applied specifically to each direction in order to find clusters of genes with monotone increasing or decreasing trends. The linear model for the δ-clustering algorithm using a reduced gene expression matrix based only on the isotonic means is given by the model in Equation (9.3) and is described in Algorithm 2.

$$y_{mi}^* = \mu + \alpha_m + \beta_i^* + r_{mi}^*. \tag{9.3}$$

Algorithm 2: Order restricted δ-clustering based on isotonic means

Input: $\boldsymbol{Y}^*$, a matrix of isotonic means, ϕ, minimum number of genes in a cluster; and λ: $0 \leq \lambda \leq 1$.

Output: Set of K clusters $\boldsymbol{Y}_k^{sub}$, $k = 1, \ldots, K$. Clusters are sub-matrices with number of rows smaller than or equal to the number of the rows of the original matrix $\boldsymbol{Y}^*$. The number of columns stays fixed, because we focus on whole profiles clustering.

Initialization: $\delta = \lambda \cdot H_P$, where H_P is the mean squared residue score of $\boldsymbol{Y}^*$.

Iteration:

1. Using the likelihood-ratio statistic, assign a direction to each gene.
2. Apply Algorithm 1 using the linear model in Equation (9.3) specifically to each direction.

9.2.2.3 Robust δ-clustering

The δ-clustering method implemented in Algorithm 1 is based on the two-way ANOVA model specified in Equation (9.2). As a consequence, the scores $H(\boldsymbol{Y})$ and $d_m(\boldsymbol{Y})$ are both computed based on residual sum of squares. As a result, similar to any least squares method, the solution of the clustering algorithm is influenced by the presence of outlying observations. In this section, we introduce the robust version of the δ-clustering method which leads to a solution that is less sensitive to outliers.

The robust δ-clustering approach is based on the median polish method (Mosteller and Tukey, 1977, Emerson and Hoaglin, 1983) and the sum of absolute residuals for the estimation of row and column effects of the cluster and the cluster membership (instead of the means and the residual sum of squares used for the δ-clustering method). The median polish algorithm is a well known robust iterative procedure in which the row and column effects are estimated by medians rather than means. Let α_m^R and β_i^R denote the row and column effects, respectively, and r_{mi}^R the residuals. We consider the model

$$y_{mi} = \mu^R + \alpha_m^R + \beta_i^R + r_{mi}^R. \tag{9.4}$$

We define the matrix MP:

$$\boldsymbol{MP} = \begin{pmatrix} r_{11}^R & \cdots & r_{1I}^R & \alpha_1^R \\ \vdots & \ddots & \vdots & \vdots \\ r_{m1}^R & \cdots & r_{mI}^R & \alpha_m^R \\ \beta_1^R & \ldots & \beta_I^R & \mu^R \end{pmatrix}.$$

The matrix MP is initialized with $r_{mi}^R = y_{mi}$ (i.e. $\alpha_m^R = \beta_i^R = \mu^R = 0$). Each iteration consists of two steps. First, for each row $m = 1, \ldots, M$, we compute the row medians and then update the matrix MP by either adding or subtracting the row medians as appropriate.

$$\begin{aligned}
\mathrm{med}_m^{row} &= \mathrm{median}(r_{m1}^R, \ldots, r_{mI}^R), \\
\mathrm{med}_{M+1}^{row} &= \mathrm{median}(\beta_1^R, \ldots, \beta_I^R), \\
r_{mi}^R &= r_{mi}^R - \mathrm{med}_m^{row}, \\
\beta_i^R &= \beta_i^R - \mathrm{med}_{M+1}^{row}, \\
\alpha_m^R &= \alpha_m^R + \mathrm{med}_m^{row}, \\
\mu^R &= \mu^R + \mathrm{med}_{M+1}^{row}
\end{aligned}$$

The first step removes the row effects from the main matrix, adds them into the α_m^R parameters and decreases the residuals. The second step applies the same procedure to the columns:

$$\begin{aligned}
\mathrm{med}_i^{col} &= \mathrm{median}(r_{1i}^R, \ldots, r_{Mi}^R), \\
\mathrm{med}_{I+1}^{col} &= \mathrm{median}(\alpha_1^R, \ldots, \alpha_M^R), \\
r_{mi}^R &= r_{mi}^R - \mathrm{med}_i^{col}, \\
\alpha_m^R &= \alpha_m^R - \mathrm{med}_{I+1}^{col}, \\
\beta_i^R &= \beta_i^R + \mathrm{med}_i^{col}, \\
\mu^R &= \mu^R + \mathrm{med}_{I+1}^{col}
\end{aligned}$$

The two steps are repeated until there is no further change in the row and column effects (Mosteller and Tukey, 1977). The resulting matrix contains all the parameters and residuals of the model represented in Equation (9.4). The row and column effects modelled by Equation (9.4) are based on medians and not the means and therefore are more robust to outliers compared to the row and column effects modelled by Equation (9.2).

We calculate a robust score for $H^R(\boldsymbol{Y})$ and $d_m^R(\boldsymbol{Y})$. In contrast with the residual sum of squares scores, discussed in Section 9.2.1, we calculate these scores using the sum of absolute residuals given by

$$H^R(\boldsymbol{Y}) = \frac{1}{MI}\sum_{m=1}^{M}\sum_{i=1}^{I}|r_{mi}^R| \qquad d_m^R(\boldsymbol{Y}) = \frac{1}{I}\sum_{i=1}^{I}|r_{mi}^R|.$$

We follow Algorithm 1 and Algorithm 2 as described above, the only change is that we use the modified residual scores, $H^R(\boldsymbol{Y})$ and $d_m^R(\boldsymbol{Y})$.

As consequence, the relative weight of r_{mi} is changed. A lower weight is put on the most extreme residuals (i.e. outlying residuals) than in the δ-clustering approach. This implies that the clusters will allow for greater deviations under the same degree of homogeneity. This property is particularly useful when the underlaying residual distribution has heavier tails than the normal distribution.

Function	Description
`monotoneDirection()`	Calculates isotonic means and classifies trends
`ORCME()`	Clusters the genes following the above mentioned Algorithm 1
`plotCluster()`	Plots the profile of genes belonging to given cluster
`plotIsomeans()`	Plots the isotonic means of the given gene
`plotLambda()`	Plots the various measures we can use to selecting the best λ parameter value
`resampleORCME()`	Applies the clustering Algorithm 1 for variety of λ values and computes various measures for λ selecting

Table 9.2: The main `ORCME` package functions.

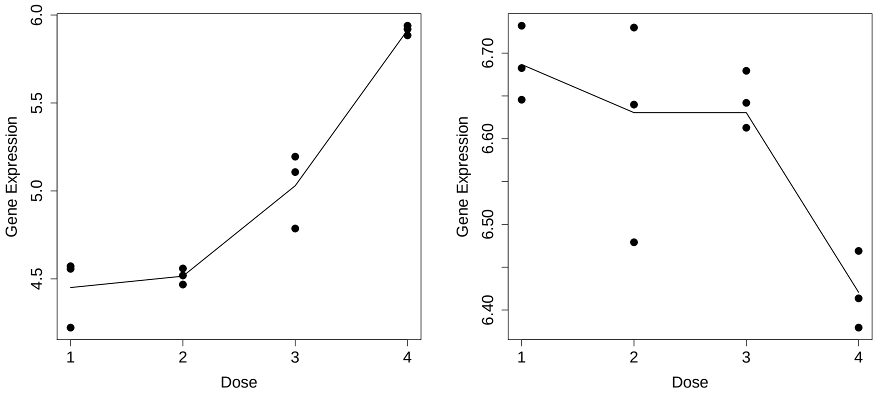
9.3 Introduction to `ORCME` package

The δ-clustering method, discussed in the Section 9.2, is implemented in the `R` package `ORCME`. The genes are clustered according to the shapes of their profiles. Primary focus is put on case of monotonicity assumption, although the clustering function can be applied in more general settings. The first stage of analysis, implementation of order restriction, is realized by function `monotoneDirection()`. It computes the isotonic means for downward and upward trends and decides which one is the most likely using the likelihood-ratio test. Isotonic means can be plotted with function `plotIsomeans()`. The δ-clustering stage is performed by the function `ORCME()` and its results can be visualized with `plotCluster()`. The homogeneity parameter λ can be estimated from the data set using the resampling procedure via function `resampleORCME()`. The results of resampling can be graphically demonstrated by function `plotLambda()`. The summary of the functions and their descriptions are presented in Table 9.2.

The `ORCME` package can be obtained from CRAN: http://cran.r-project.org/web/packages/ORCME/index.html. The `ORCME` package requires the package `Iso` (Lin *et al.*, 2013).

9.3.1 Example 1: δ-clustering for dose-response data

In this section we illustrate the use of the package ORCME on the HESCA data set that is described in Section 1.2. Note that the primary interest is to cluster genes with monotone gene profiles and therefore, similar to Lin *et al.* (2009), we prefer to perform an inference step before the actual clustering. Similar approach for order restricted, but not monotone profiles, is discussed by Peddada *et al.* (2005). As mentioned above, the initial step for dose-response microarray data is performed by applying likelihood-ratio test to establish a dose-response relationship under order restricted constraints. Non-significant genes are excluded and the significant genes are assigned to the monotone direction with higher likelihood. In total, the null hypothesis was rejected for 2,910 out of the 16,998 genes that were tested, with 1,321 upwards and 1,589 downwards regulated genes. Examples of significantly increasing and decreasing trends are shown on Figures 9.1a and 9.1b, respectively. Note that subset of this data set is used as example data in the package.



(a) Upward trend.

(b) Downward trend.

Figure 9.1: Examples of two significant genes.

The first step in ORCME package is typically to distinguish between upward and downward directions of significant genes. The function `monotoneDirection()` can be used to identify the direction of the trend. The applying of `monotoneDirection()` can take several minutes for large data sets.

```
R> library("ORCME")
R> dim(geneData)
[1] 2910    12
```

```
R> geneData[1:5,1:6]
           X1     X1.1     X1.2       X2     X2.1     X2.2
[1,] 6.923109 7.024719 7.170328 7.219297 7.076908 7.404949
[2,] 6.695870 6.687039 6.652153 6.503670 6.387794 6.698711
[3,] 3.976558 4.016001 4.631135 4.335205 4.264335 4.679793
[4,] 5.379032 4.961081 5.691166 5.193203 5.231240 5.496361
[5,] 6.097025 6.263939 6.217385 6.551656 6.632323 6.335757
R> doseData <- c(1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4)
R> dirData <- monotoneDirection(geneData = geneData, doseData = doseData)
```

Secondly, after the determination of the trend direction, we create the R objects for genes with upward and downward trends. The output contains list of monotone trend direction for each gene, isotonic means for each gene and lists of isotonic means and observed values for genes classified as upwards and downwards separately. Then, the function plotIsomeans() can be used to produce gene-specific profile plot as is shown in Figure 9.1.

```
R> Direction <- dirData$direction
R> Direction[1:5]
[1]  "up" "up" "up" "dn" "up"

R> incData <- as.data.frame(dirData$incData)
R> dim(incData)
[1] 1321    4
R> incData[1:5,]
        V1       V2       V3       V4
1 7.039385 7.233718 7.402824 7.795044
2 6.604206 6.604206 6.968700 8.992689
3 4.207898 4.414689 4.414689 5.006698
4 6.192783 6.396748 6.396748 7.029999
5 3.468541 3.468541 4.300872 9.086498

R> decData <- as.data.frame(dirData$decData)
R> decData[1:5,]
        V1       V2       V3       V4
1 5.343760 5.306935 4.982664 4.083754
2 7.760716 7.462762 7.199676 6.812242
3 5.963389 5.963389 5.705701 5.198672
```

```
4 7.000747 6.999338 6.815059 6.544177
5 7.155324 7.155324 6.987014 6.719862

R> obsIncData <- as.data.frame(dirData$obsincData)
R> obsIncData[1:5,1:6]
        X1     X1.1     X1.2       X2     X2.1     X2.2
1 6.923109 7.024719 7.170328 7.219297 7.076908 7.404949
2 6.695870 6.687039 6.652153 6.503670 6.387794 6.698711
3 3.976558 4.016001 4.631135 4.335205 4.264335 4.679793
4 6.097025 6.263939 6.217385 6.551656 6.632323 6.335757
5 3.411178 3.776434 3.346341 3.436557 3.556874 3.283860

R> obsDecData <- as.data.frame(dirData$obsdecData)
R> obsDecData[1:5,1:6]
        X1     X1.1     X1.2       X2     X2.1     X2.2
1 5.379032 4.961081 5.691166 5.193203 5.231240 5.496361
2 7.422136 7.674339 8.185674 7.330481 7.685623 7.372183
3 5.743115 5.910513 6.090424 5.753453 6.298518 5.984308
4 7.074718 7.052492 6.875032 6.778505 7.108147 7.111361
5 7.049960 7.091273 7.105668 7.175221 7.150928 7.358895

R> isoMeans <- as.data.frame(dirData$arrayMean)
R> isoMeans[1:5,]
        V1       V2       V3       V4
1 7.039385 7.233718 7.402824 7.795044
2 6.604206 6.604206 6.968700 8.992689
3 4.207898 4.414689 4.414689 5.006698
4 5.343760 5.306935 4.982664 4.083754
5 6.192783 6.396748 6.396748 7.029999

R> plotIsomeans(monoData=incData, obsData=obsIncData,
+ doseData=doseData, geneIndex=10)
```

The main function for clustering is ORCME(). Based on the penalized within cluster sum of squares (which will be discussed in Section 9.4), $\lambda = 0.15$ is chosen as the optimum choice of λ for clustering the upward monotone genes (algorithm is described in Section 9.4). In our example, for genes with upward trends we use following code (be aware that computation can take several minutes) and the output is a matrix of genes in rows and found clusters in columns. Because one gene can be present only in one cluster, there is only one TRUE value in each row.

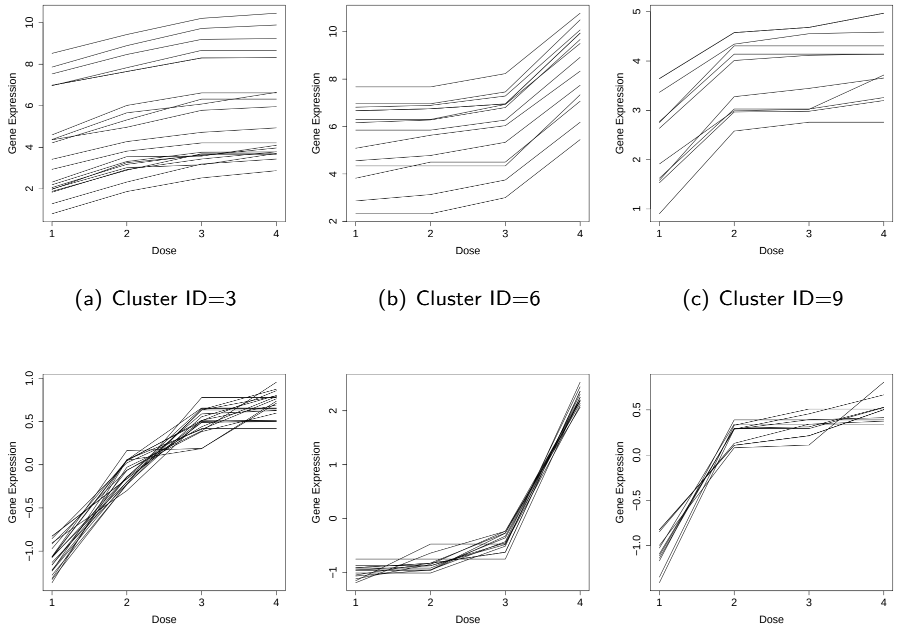
```
R> ORCMEoutput <- ORCME(DRdata = incData, lambda = 0.15, phi = 2)
R> dim(ORCMEoutput)
[1] 1321   27

R> ORCMEoutput[1:5,1:6]
      V1    V2    V3    V4    V5    V6
g1 FALSE  TRUE FALSE FALSE FALSE FALSE
g2  TRUE FALSE FALSE FALSE FALSE FALSE
g3 FALSE  TRUE FALSE FALSE FALSE FALSE
g4  TRUE FALSE FALSE FALSE FALSE FALSE
g5  TRUE FALSE FALSE FALSE FALSE FALSE
```

The δ-clustering method with $\lambda = 0.15$ results in 27 clusters for 1,321 upward monotone genes. The first cluster contains 1,051 genes and the last ones contain only two genes (this number was set as the minimal cluster size). The large size of the first cluster is an inherent feature of the δ-clustering method. The first clusters from the δ-clustering method often contain genes that are less expressed and less variable than those in the later clusters. Figure 9.2 presents examples of clusters with upward monotone profiles. The upper panel shows the raw gene expression values and the lower panel show gene expression values centered around gene specific means. Figure 9.2 clearly shows that genes within a cluster have coherence in terms of similarities between their expression values and trends. The function `plotCluster()` produces the isotonic mean profiles for a specific cluster. The option `zeroMean = TRUE` centers the gene profiles around the gene-specific means, as shown in the lower panels of Figure 9.2.

```
R> plotCluster(DRdata = incData, doseData = doseData,
+ ORCMEoutput = ORCMEoutput, clusterID = 3, zeroMean = FALSE,
+ xlabel = "Dose", ylabel = "Gene Expression")
```

The penalized within cluster sum of squares score suggests $\lambda = 0.35$ as the optimum choice of λ for the downward monotone genes. The application of the δ-clustering method results in 19 clusters for the 1,589 downward monotone genes. The first cluster contains 1,433 genes and the last cluster contains two genes. Figure 9.3 presents examples of clusters with downward monotone profiles. Similar to the clustering of the upward monotone genes shown on Figure 9.2, the clusters contain genes with coherent values. However, we



(a) Cluster ID=3

(b) Cluster ID=6

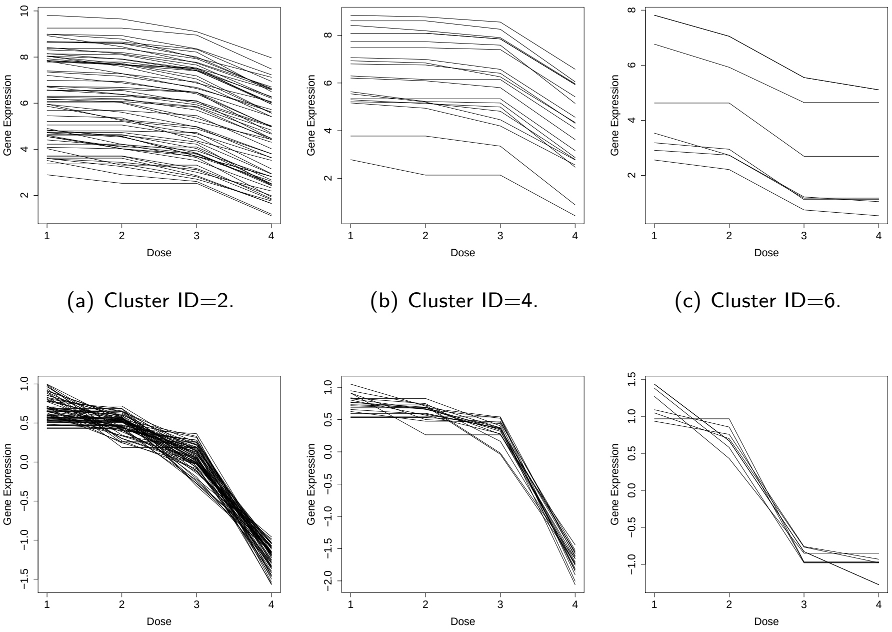
(c) Cluster ID=9

(d) Cluster ID=3 (centered)

(e) Cluster ID=6 (centered)

(f) Cluster ID=9 (centered)

Figure 9.2: Examples of clusters from upward monotone genes. Top panels: Gene expression profiles. Bottom panels: Centered gene expression profiles.



(a) Cluster ID=2.

(b) Cluster ID=4.

(c) Cluster ID=6.

(d) Cluster ID=2 (centered).

(e) Cluster ID=4 (centered).

(f) Cluster ID=6 (centered).

Figure 9.3: Examples of clusters with downward monotone profiles. Top panels: Gene expression profiles. Bottom panels: Centered gene expression profiles.

can notice situations where few members of a cluster show different dose-response trends, although such deviations typically occur in one of the four doses in the experiment.

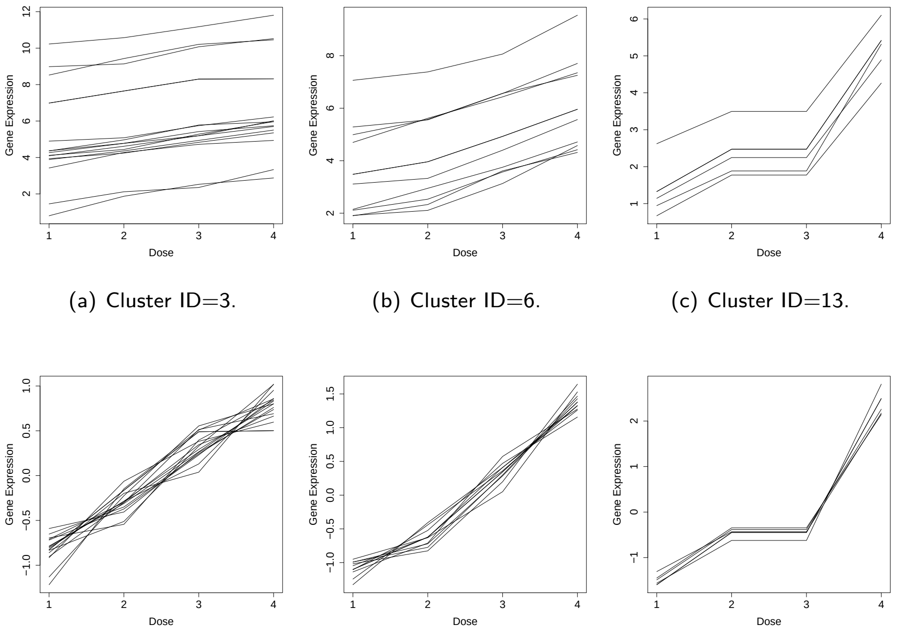
The robust version of the algorithm can be called with change of appropriate option and results plotted as demonstrated before:

```
R> ORCMEoutputRobust <- ORCME(DRdata = incData, lambda = 0.55,
+ phi = 2, robust=TRUE)
R> plotCluster(DRdata = incData, doseData = doseData,
+ ORCMEoutput = ORCMEoutputRobust, clusterID = 4, zeroMean = FALSE,
+ xlabel = "Dose", ylabel = "Gene Expression")
```

The value of $\lambda = 0.55$ was selected to achieve an optimal balance between the within cluster variability and the number of clusters. The robust algorithm was applied to the set of 1,321 upward monotone genes resulted in 26 clusters (compared to 27 when ANOVA was used). The first cluster contains 1,077 genes and several of the smallest clusters containing only two genes (this number was set as the minimal cluster size). Figure 9.4 presents three examples of clusters with upward monotone profiles. Compared to the non-robust version of the algorithm (Figure 9.2), the clusters contain genes with higher within cluster variability. This is due to the fact that a lower penalty is given to extreme outliers. Therefore, the method allows for higher dissimilarities within the cluster.

Finally, the results obtained from the δ-clustering method using the ORCME() package for 1,321 upward monotone genes were compared with hierarchical clustering results in order to understand the benefits of the δ-clustering method. The cosine similarity method (Salton, 1988) was used to measure distance of the profiles. The cosine similarity takes into account both scale and shape of the mean profile, a property that it shares with the δ-clustering method. The hierarchical clustering with Ward's linkage was applied and resulting dendrogram was cut in order to obtain same number of clusters as produced by δ-clustering (27 clusters). The following R code used to performed the hierarchical clustering:

```
R> Y <- as.matrix(incData)
R> X <- as.matrix(Y/sqrt(rowSums(Y^2)))
R> cosD <- as.dist(1 - crossprod(t(X)))
R> out <- hclust(cosD, method = "ward.D2")
R> out <- cutree(out, ncol(ORCMEoutput))
R> id <- names(sort(table(out), decreasing = TRUE))
R> center <- function(X) t(X - rowMeans(X))
R> matplot(center(incData[which(out == id[1]),]), type = "l", col="black",
+  lty=1, axes=FALSE, xlab="Dose", cex.lab=1.5, ylab="Gene expression")
```



(a) Cluster ID=3. (b) Cluster ID=6. (c) Cluster ID=13.

(d) Cluster ID=3 (centered). (e) Cluster ID=6 (centered). (f) Cluster ID=13 (centered).

Figure 9.4: Robust δ-clustering. Examples of clusters with upward monotone profiles. Top panels: Gene expression profiles. Bottom panels: Centered gene expression profiles.

```
R> axis(2, cex.axis=1.5)
R> axis(1, at=seq(1:4), label=seq(1:4), cex.axis=1.5)
```

The first four clusters for both cosine similarity based hierarchical clustering and δ-clustering are shown in Figure 9.5. The results seem to be similar, but the clusters produced by δ-clustering seem to be more homogeneous (this difference is the most profound in the fourth cluster, the panels in Figure 9.5). The δ-clustering is designed to find clusters with similar dose-response shapes. The value of homogeneity parameter prevents appearance of such clusters as on top right panel of Figure 9.5, as found by hierarchical clustering method. In the context of dose-response experiments, the advantage of the δ-clustering method is that it relies directly on the cluster-specific parameter estimates for the monotone dose effects, β_i, and gene effects, α_m (i.e. Equation 9.3 and Equation 9.4) which are of primary interest for dose-response transcriptomics experiments. An additional advantage of the δ-clustering is the use of correction mechanism (discussed in Section 9.2.2 as node addition step) that adds additional genes into clusters that were found in the previous step of the algorithm.

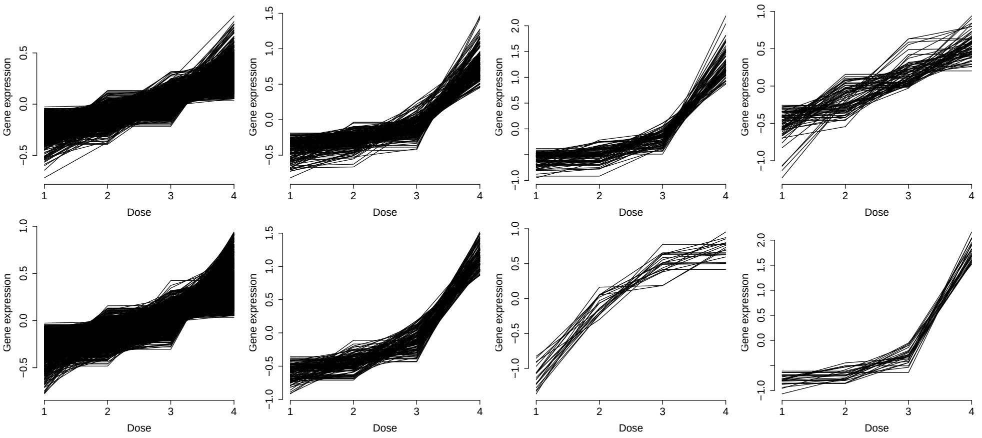


Figure 9.5: Comparison between cosine similarity based hierarchical clustering (top panels) and δ-clustering (bottom panels) results.

9.4 Choice of clustering parameter λ

How to estimate the optimal number of clusters is a major challenge in cluster analysis. In most cases, the quality of such estimate determines the quality of the resulting clusters. While the number of clusters is not required for the δ-clustering method, the optimum choice for λ and ϕ is unknown and may be data dependent. We suggest that ϕ is fixed at a pre-specified value and the choice of lambda is explored based on the data (for details see Section 9.5). The possible choice for λ can be based on the within cluster sum of squares, which can be computed for λ in the range of zero to one. Let us assume that for a specific value of λ the δ-clustering method results in $n(\lambda)$ clusters denoted $C_1, \ldots, C_{n(\lambda)}$. Let $R(\lambda)$ denote the within cluster sum of squares for this value of λ, then

$$R(\lambda) = \sum_{q=1}^{n(\lambda)} \sum_{m \in C_q} \sum_{i=1}^{I} (y^*_{mi} - \mu_q - \alpha_m - \beta^*_{iq})^2,$$

where μ_q and β^*_{iq} are cluster specific parameters. Recall that M is the number of genes to be clustered. The range for $n(\lambda)$ lies between one and the number of genes, i.e, $1 \leq n(\lambda) \leq M$. When $\lambda = 1$, $n(\lambda) = 1$ and $n(\lambda) \leq M$ for $\lambda = 0$. Since $R(\lambda)$ is a decreasing function of $n(\lambda)$ and an increasing function of λ, $R(\lambda)$ will be minimal when $n(\lambda) = M$ and maximal when $n(\lambda) = 1$. Note that when $n(\lambda) = 1$, the within cluster sum of squares equals the total sum of squares for the gene expression matrix. Our aim is to find the value of λ when taking the trade-off between the within cluster sum of squares and the number of resulting clusters into account. This criterion is referred to as *penalized within cluster sum of squares (pWSS)* and it is defined as

$$pWSS(\lambda) = R(\lambda) + 2n(\lambda).$$

Following Tibshirani *et al.* (2001), other criteria for traditional clustering methods can be considered as well. We can modify the Calinski and Harabasz (1974) index as

$$CH(\lambda) = \frac{B(\lambda)/n(\lambda)}{W(\lambda)/(M - n(\lambda))},$$

where $B(\lambda)$ and $W(\lambda)$ are the between cluster sum of squares and within clusters sum of squares, respectively. While the within cluster sum of squares is expected to increase with increasing λ, the between cluster sum of squares is expected to decrease. Another criterion is the Hartigan and Wong (1979) index, which is also modified as

$$H(\lambda) = \left[\frac{W(\lambda_\ell)}{W(\lambda_{\ell+1})} - 1\right] \cdot \frac{1}{M - n\left(\lambda_{\ell+1}\right)},$$

where ℓ is an index for the unique value of λ. The original definition for the H index is based on the sequential increase in number of clusters. For our proposal, this is not the case, as more than one value of λ may result in the same number of clusters. However, the criterion can still be used to investigate the gain in within cluster sum of squares when moving from a lower value of λ to an adjacent higher value.

For the robust version of δ-clustering, we use a robust version for the within cluster sum of squares:

$$R^R(\lambda) = \sum_{q=1}^{n(\lambda)} \sum_{m \in C_q} \sum_{i=1}^{I} |y^*_{mi} - \mu_q - \alpha_m - \beta^*_{iq}|,$$

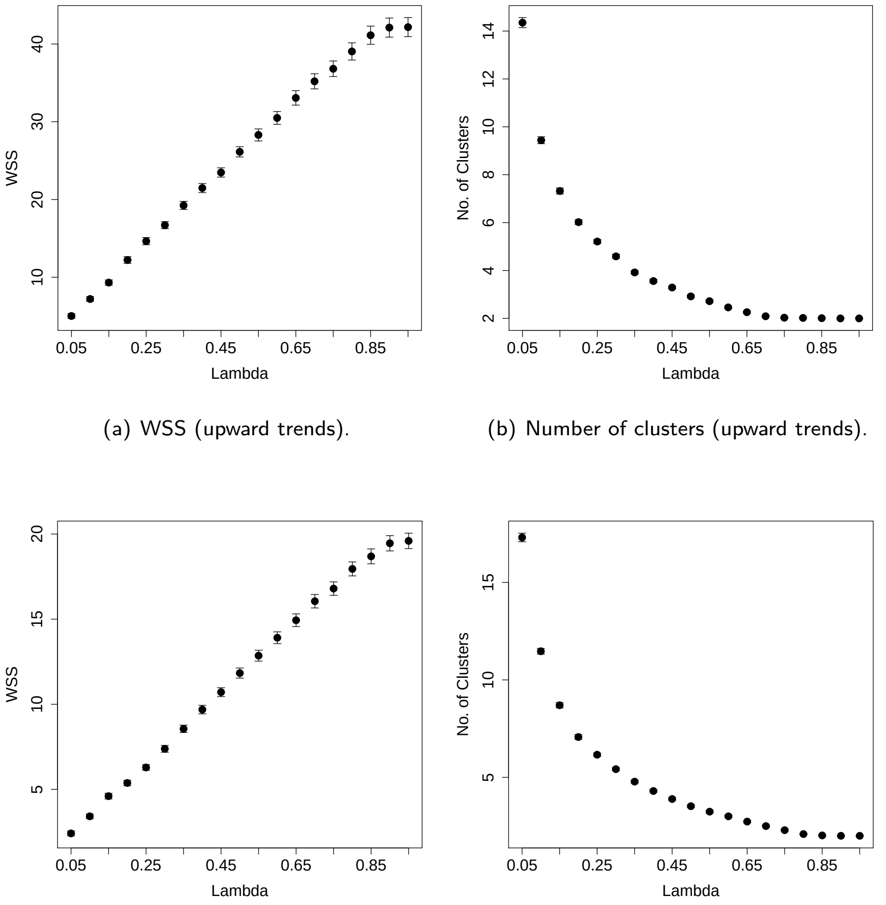
where the row and column effects are computed using median polish algorithm. Otherwise, the procedure follows the same workflow. Note that the robust algorithm would generally lead to lower number of clusters if *pWSS* is used. The use of the absolute value will result in lower values of within cluster variability (compared to ANOVA method), therefore the penalty term based on $n(\lambda)$ will be relatively more influential compared to ANOVA case.

9.4.1 Example 2: The choice of the clustering parameter

9.4.1.1 The trade-off between clustering parameter λ and the number of clusters

The relative proportion (λ) of the mean squared residue score of the monotonised gene expression matrix is proposed as a clustering parameter for the δ-clustering method. Though λ is bounded between zero and one, the choice of the optimum value of λ is unknown. Similar to the resampling approach for random forest (Breiman, 1996) and ABC learning (Amaratunga *et al.*, 2008), we propose to generate 100 resampled data sets, with each data set containing 100 genes randomly sampled with replacement from the reduced expression data. Reduced expression data means that the clustering is typically applied after initial filtering of genes. For each of the resampled data sets, the δ-clustering method is applied based on a set of values of λ ranging from 0.05 to 0.95. Note that the minimum number of genes in a cluster is fixed at two. The resampling is done using the function `resampleORCME()` and typically it can take several minutes. The output consist of within cluster sum of squares, total sum of squares and number of clusters for particular lambda.

```
R> lambdaVector <- seq(0.05, 0.95, 0.05)
R> lambdaChoiceOutput <- resampleORCME(clusteringData = incData,
+ lambdaVector = lambdaVector)
R> lambdaChoiceOutput[[1]][1:10,]
      lambda       WSS      TSS nc
 [1,]   0.05  3.358893 43.61771 16
 [2,]   0.10  4.448680 43.61771 10
 [3,]   0.15 10.310378 43.61771  6
 [4,]   0.20  8.090213 43.61771  6
 [5,]   0.25 13.516462 43.61771  5
 [6,]   0.30 12.772803 43.61771  5
```



(a) WSS (upward trends).

(b) Number of clusters (upward trends).

(c) WSS (downwards trends).

(d) Number of clusters (downward trends).

Figure 9.6: Within clusters sum of squares and the number of clusters as a function of λ.

```
 [7,]   0.35 14.942951 43.61771  5
 [8,]   0.40 19.533244 43.61771  4
 [9,]   0.45 22.261206 43.61771  4
[10,]   0.50 20.122818 43.61771  4
```

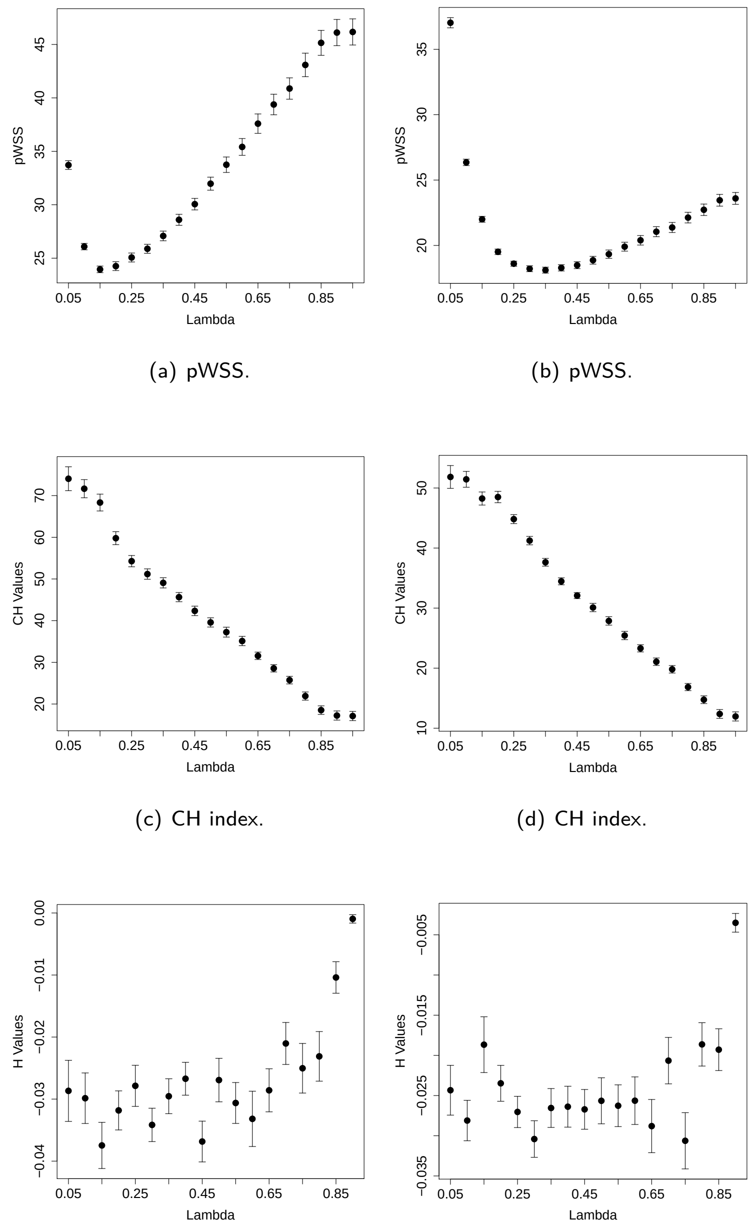
Figure 9.6 shows the relationship between the within cluster sum of squares, the number of resulting clusters, and λ. Panels 9.6a and 9.6c show the relationship between the within cluster sum of squares and λ for the upward and downward monotone genes, respectively. Panels 9.6b and 9.6d show the relationship between the number of resulting clusters and λ for the upward and downward monotone genes, respectively. The within cluster sum of squares increases with an increase in λ, while the number of clusters decreases with an increase in λ. It shows that a trade-off between the within cluster sum of squares and number of clusters may be a criterion to choose an optimal λ. Diagnostic plots were produced using the function `plotLambda()`.

```
R> plotLambda(lambdaChoiceOutput, output = "wss")
R> plotLambda(lambdaChoiceOutput, output = "ncluster")
```

9.4.1.2 The choice of clustering parameter λ

The trade-off between the within cluster sum of squares and number of clusters is visualized in Figure 9.7. Penalized within cluster sum of squares (*pWSS*) is presented in panels 9.7a and 9.7b for upward and downward monotone genes, respectively. The *pWSS* reaches a minimum at $\lambda = 0.15$ for the upward monotone genes and at $\lambda = 0.35$ for the downward monotone genes. Panels 9.7c and 9.7d show the relationship between the *CH* values and λ for upward and downward monotone genes, respectively. The maximum value of *CH* is reached at $\lambda = 0.05$ for both the upward and downward monotone genes. It appears for our case study that the *CH* index is not an informative criterion. It favors the λ value which results in the highest number of clusters. Panels 9.7e and 9.7f present the relationship between the *H* value and λ for the upward and downward monotone genes, respectively. The *H* values do not show a smooth pattern as observed from the *pWSS*. However, it reaches its minimum at $\lambda = 0.15$ for the upward monotone genes and at $\lambda = 0.75$ (note that second lowest value is at $\lambda = 0.30$) for the downward monotone genes. Graphical output can be produced using the function `plotLambda()`. The option `output="..."` determines which index will be plotted.

```
R> plotLambda(lambdaChoiceOutput, output="pwss")
R> plotLambda(lambdaChoiceOutput, output="ch")
R> plotLambda(lambdaChoiceOutput, output="h")
```



(a) pWSS.

(b) pWSS.

(c) CH index.

(d) CH index.

(e) H index.

(f) H index.

Figure 9.7: The choice of λ. Left panels: Upward trends. Right panels: Downward trends.

9.5 Discussion

Gene clustering is one of the topics of interest in the analysis of the dose-response microarray experiments. The aim is to find clusters of genes with similar dose-response relationships under an increasing dose of a therapeutic compound. In the `ORCME` package, the δ-clustering method is proposed for the clustering of dose-response microarray data. The method is motivated by the δ-biclustering method proposed by Cheng and Church (2000), where they define a bicluster as a subset of genes and a subset of conditions with a "high similarity score" using the mean squared residue score. For the δ-clustering method, the δ value is modified to be data dependent. It is expressed as a relative proportion (λ) of the mean squared error of the gene expression matrix (as if all genes are treated as if they belong to a single cluster). The method shares some features with standard clustering methods (it partitions genes into non-overlapping groups), but it also benefits from the local structures of biclustering methods.

The δ-clustering procedure should be applied to a reduced expression matrix obtained after initial (inference-based) filtering to keep only the significant genes in the cluster analysis. The optimum choice of λ is explored with penalized within cluster sum of squares, which offers a trade-off between the goodness-of-fit and the complexity of the resulting clusters, for λ values ranging from zero to one. The goodness-of-fit is captured by the within cluster sum of squares and the complexity is captured by the number of clusters. Note that the within cluster sum of squares increases with an increase in λ and the number of clusters decreases with an increase in λ. No optimization tool exists for selecting an optimal value of ϕ. It is suggested that ϕ is fixed by the user bearing in mind that ϕ can be interpreted as the smallest cluster size. ϕ is not related to the composition of the cluster, but has a practical interpretation as the smallest cluster size of interest. We expect that in general ϕ would be set to two in order to identify as many clusters as possible. In contrast, interpretation of λ as the homogeneity parameter is more abstract and it's specification is therefore more difficult.

The method and package were introduced within the framework of order restricted microarray experiments with special focus on the monotonicity assumption. As mentioned earlier, the methodology can be applied to any other type of order restriction. Separating the two stages allows the user to compute the means under any type of restrictions (such as umbrella profiles), or without any restriction at all. Then, the function `ORCME` can be applied without any change as described above. Moreover, the application is not necessarily limited to microarray experiments. The δ-clustering method can be applied to any data where whole-profile clustering is of interest. It can be other biological data, from related fields (such as metabolomics, proteomics and RNAseq) or in the broader context (NMR data, immunological data or data from public health studies). In general, the method can be applied in any situation where the aim is to cluster the observations according to the behaviour of some response on ordered categories. As such, other areas for potential application include, but are not limited to, environmental studies and financial studies.

Other methods for the profile clustering are ORICC (Liu *et al.*, 2009) and the method implemented in the `ORIOGEN` package. ORICC is implemented in the `R` package `ORIClust` and is based on the order restricted information criterion which implements the Akaike information criterion (Akaike, 1974) idea in the case of monotonicity. The `Java` software `ORIOGEN` is based on the idea of inference-based clustering using bootstrap resampling described in Peddada *et al.* (2003). Note that both `ORIClust` and `ORIOGEN` cluster genes with order restricted profiles, but not necessarily monotone profiles, such as umbrella profiles. They are not designed to distinguish particular profiles within the monotone profiles, instead they pool them in one cluster. In contrast, `ORCME` is a very suitable clustering algorithm when the monotone profiles are of primary interest and subcategories need to be distinguished.

Chapter 10

A Community Based Software development: The IsoGeneGUI Package

10.1 Introduction

Modelling dose-response relationship plays an important role in the drug discovery process in the pharmaceutical industry. Typical responses are efficacy or toxicity measures that are modelled with the aim of identifying the dose that is simultaneously efficacious and safe (Pinheiro *et al.*, 2006). The recent development of microarray technology introduced gene expression level as an additional important outcome related to dose. Genes, for which the expression level changes over the dose of the experimental drug, are of interest, since they provide insight into efficacy, toxicity and many other phenotypes. Order restriction is often assumed in the dose-response modelling, usually in terms of monotone trend (Lin *et al.*, 2012d). The restriction is a consequence of the assumption that higher dose levels induce stronger effects in the response (either increasing or decreasing). However, order restriction can also be related to umbrella profiles. In such a case, monotonicity is assumed up to a certain dose level and the direction of the dose-response relationship changes thereafter (Bretz and Hothorn, 2003).

Order restricted analysis received a lot of attention in previous years and several R (R Core Team, 2013) packages were developed for this purpose. Specifically, the R packages IsoGene (Lin *et al.*, 2013 and Pramana *et al.*, 2010) and orQA (Klinglmueller, 2010)

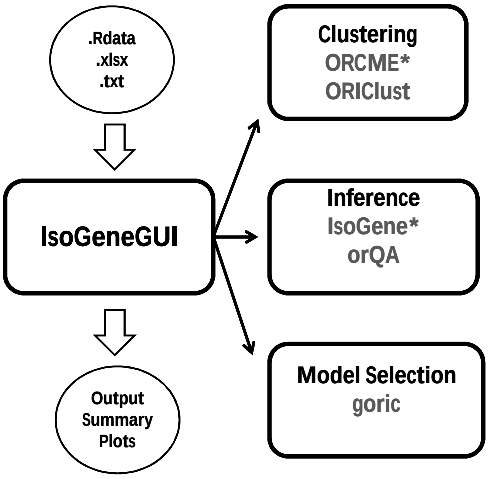


Figure 10.1: The general structure of the `IsoGeneGUI` package. The ones notated by asterisk were developed and maintained by same research group as the `IsoGeneGUI`, all the remaining are work of different scientific groups.

were developed for inference, `goric` (Gerhard and Kuiper, 2012) for model selection, and `ORCME` (Kasim *et al.*, 2014) and `ORIClust` (Liu *et al.*, 2012) were developed for order restricted clustering of genes.

Inference consists of testing a null hypothesis of no dose-response relationship, against an ordered alternative. Multiplicity correction needs to be applied due to the large number of tests. The model selection framework quantifies the expected relative distance of a given model to the true underlying model in order to select the best model among a set of candidate models. The model selection approach is basis for the identification of the minimal effective dose or lowest-observed-adverse-effect level (Kuiper *et al.*, 2014). Order restricted clustering is a data analysis approach which aims to form subsets of genes with similar expression profiles. It is very useful when reference genes are available and the aim of the analysis is to identify genes that behave in a similar way to the reference genes. All the different methods were scattered across multiple specialized packages. The `IsoGeneGUI` package is an envelope package in which all the methods are available together in user friendly framework, allowing to explore the gene expression data set with collection of state-of-the-art tools. The overview of the package structure is schematically shown in Figure 10.1.

Not all scientists performing microarray experiment analysis are necessarily educated in using R. Hence, the package IsoGeneGUI (Pramana *et al.*, 2012b) was originally created as a graphical user interface extension of the IsoGene package. The large number of IsoGeneGUI package downloads from the BioConductor (Gentleman *et al.*, 2004) repository suggests that there is a demand for GUI data analysis tools for inference, model selection, estimation and order restricted clustering. Therefore, the IsoGeneGUI package was extended to embrace all currently available tools in one package. In addition to the data analysis tools for inference, model selection, clustering and estimation the package contains many tools for exporting results, their visualization and easy handling of produced figures. Therefore, IsoGeneGUI provides the most complete and simultaneously user friendly data analysis tool dealing with order restricted microarray experiments that is currently available in R.

In this chapter, we provide a brief introduction to the package, both underlying methodology and its particular implementation. The general principles of GUI are explained in Section 10.2. Methods for estimation, inference, clustering and model selection available in IsoGeneGUI package are introduced in Section 10.3. The structure of the package is described in Section 10.4 and details about implementation of the methods are given in Section 10.5, accompanied by multiple figures illustrating the GUI environment. Final summarization in Section 10.6 concludes the chapter.

10.2 GUI packages

The IsoGeneGUI represents the connection of two principles in modern software development in R: graphical user interface and envelope packages. Both of the principles aim to improve the experience with R and provide the user with friendly and clear tool. The general knowledge of the R software mechanisms is still necessary to use the package properly, but large amount of details related to coding and technical part of R are not relevant for the user of the GUI which significantly speed up learning process and simplifies the analysis workflow.

The main advantage of GUI is intuitive specification of parameters and running the functions with button clicks rather than typing commands. Not only it saves the user from typos and programming mistakes, but it also allows him/her to use the functions without knowing their exact name and command. The main disadvantage is the necessary simplification of the analysis. Since most of the GUI originates from the usual command line-based R packages, not all the functions can be easily converted into the GUI environment. Typical example are functions that require prior specification of more complex object as list or matrix to be used as input. The construction of such objects would be overly complicated in windows-based environment and conflicting in targeted clarity of the package. Therefore, some options of the original functions of converted package may be omitted. Secondly, the simplification is related to omission of outputted code. User does not see how his/her instructions were translated into the R code and therefore cannot modify the commands for different purposes. However, both of these simplifications are exactly following line of development of GUI: to keep the analysis as simple and as user-friendly as possible. The users that require full flexibility of the original packages or need to modify can always access the original packages through command line and perform the analysis via basic R environment.

Multiple R packages are created as ensemble of the methods, providing whole range of analysis options. Envelope packages are further extension, providing whole range of packages within one GUI environment, without user necessarily knowing he needs to apply multiple packages to preform analysis of interest. The Comprehensive R Archive Network (CRAN, Hornik, 2012) and Bioconductor, two main R repositories, currently contain more than 6,000 packages. The open source nature of R project does not allow for curation of all added packages, so there are multiple packages dealing with same type of data or analysis of interest, often taking slightly different perspectives, methodology and interpretation of results. Searching for all the possibilities can be challenging for unexperienced user, as well as evaluation of the quality of the particular package. Therefore, envelope packages are developed. They combine several packages into one entity providing wide range of methods and guaranteeing at least some degree of peer review of the methodology and programming part.

Synergy of these two principles creates the envelope GUI packages, such as IsoGeneGUI package. The final package then contains most of the available methodology dealing with the topic of interest, together with unified framework for evaluation, interpretation, saving and visualization of results, everything in user friendly window-based environment. The authors of the most of the particular original R packages were involved in the late stage of development in order to check the performance of their package within the IsoGeneGUI package and to advice on the exact implementation of the methods.

10.3 Order restricted analysis of continuous data

The functionality of the package can be divided into three areas: inference, clustering and determination of the minimum effective dose. Additional tasks, such as estimation of dose-effects, model selection and model averaging can be implemented within the package as well.

The main goal of the inference framework is to test the null model of no dose effect against an ordered alternative. Several test statistics for order restricted problems were developed over the last few decades. In the package, the following methods are available: likelihood-ratio test (LRT, Barlow *et al.*, 1972), Williams' test statistic (Williams, 1971), Marcus' statistic (Marcus, 1976), M statistic (Hu *et al.*, 2005) and modified M statistic (Lin *et al.*, 2007). Detailed elaboration about the methods, their usage and advantages and disadvantages can be found in Lin *et al.* (2007) and Lin *et al.* (2012d). The distribution of some of the test statistics cannot be derived analytically. Therefore, resampling based inference is implemented to approximate distribution of test statistics under the null model (Westfall and Young, 1993 and Ge *et al.*, 2003). When the tests are performed for a large number of genes, the multiplicity adjustment is necessary. Family Wise Error Rate (FWER) can be controlled by Bonferroni (Bonferroni, 1936), Holm (Holm, 1979), Hochberg (Hochberg and Benjamini, 1990) or Šidák single-step and step-down (Šidák, 1971) procedures. Alternatively, False Discovery Rate (FDR) can be controlled with the Benjamini-Hochberg (BH, Benjamini and Hochberg, 1995) or Benjamini-Yekutieli (BY, Benjamini and Yekutieli, 2001) procedures. A common issue in gene expression inference is the presence of genes with relatively low variance that induce large values of the test statistics under consideration, although the magnitude of the effect is negligible. Formally, the genes are declared statistically significant, but from a biological point of view, these genes will not be further investigated due to small fold change. Significance Analysis of Microarrays (SAM, Tusher *et al.*, 2001) was proposed as a solution for this issue by inflating the standard error.

The `IsoGeneGUI` package provides two clustering approaches based on algorithms that incorporate order restrictions. The `ORCME` package implements the δ-clustering algorithm (Kasim *et al.*, 2012) which is based on the δ-biclustering algorithm proposed by Cheng and Church (2000). It is described in detail in Chapter 9. It is applied to isotonic means and hence ignores the within dose variability and uncertainty about the mean estimation. Therefore, it is advised that the algorithm is applied either to a filtered data set (i.e. genes with fold change higher than given threshold) or on the genes showing significant dose-response profile (i.e. after the inference step).

The `ORIClust` package implements the one or two-stage Order Restricted Information Criterion Clustering algorithm (ORICC, Liu *et al.*, 2009, Lin *et al.*, 2012c) which is based on an information criterion that takes into account order restrictions. The filtering step can be addressed within the algorithm itself. The ORICC algorithm considers different type of dose-response profiles, such as monotone profiles and umbrella profiles, that can be used for clustering. Umbrella profiles assumes that the monotonicity holds up to a certain dose and then the trend changes the direction. Practical example, when such profiles are suitable, is overdosing with the drug, changing beneficial effect to the harmful one. In contrast to the clustering approach implemented in the δ-clustering method, the

ORICC algorithm pulls together all monotone profiles. Hence, it is not suitable for the separation of non-decreasing monotone profiles with a true zero effect at some dose levels (i.e. some dose-specific means are equal) from strictly increasing profiles. This is the main difference between these two clustering algorithms proposed by Liu *et al.* (2009) and Lin *et al.* (2012c) and the reason why they are both needed to provide a complete toolbox for an order restricted analysis of microarray data.

A model selection based method is implemented in the package `goric` using Generalized Order Restricted Information Criterion (GORIC, Kuiper *et al.*, 2011). Similar to the ORIC (Anraku, 1999) algorithm, the GORIC method incorporates the information about the order constraints when calculating the information criteria. The minimum effective dose can be selected based on GORIC weights (Kuiper *et al.*, 2014) that can be interpreted as posterior model probabilities (Lin *et al.*, 2012c). Details about GORIC procedure and model selection in general can be found in Chapter 4.

10.4 The structure of the package

The package `IsoGeneGUI` encompasses all the methods mentioned in previous section. The summary is given in Table 10.1. The GUI was build using `Tcl/Tk` environment.

Package	Analysis type	Reference
`IsoGene`	Inference	Lin *et al.* (2012d)
`orQA`	Inference	Klinglmueller (2010)
`ORCME`	Clustering	Kasim *et al.* (2014)
`ORIClust`	Clustering	Liu *et al.* (2012)
`goric`	Model selection	Gerhard and Kuiper (2012)

Table 10.1: *Packages for the analysis of order-restricted dose-response gene expression data available on CRAN.*

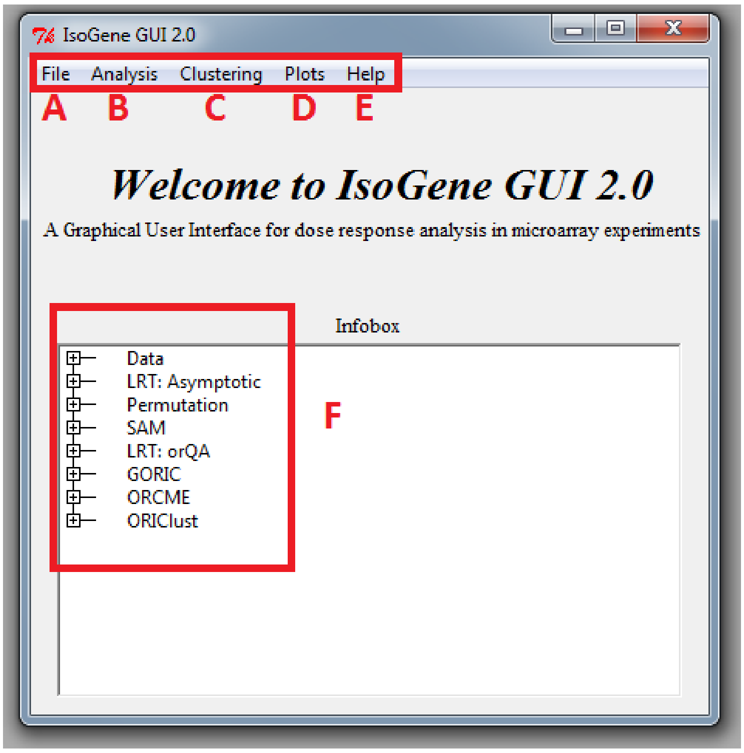
The `IsoGeneGUI` is freely available from Bioconductor repository. It can be downloaded and run from `R` with commands:

```
source("http://bioconductor.org/biocLite.R")
biocLite("IsoGeneGUI")
library(IsoGeneGUI)
IsoGeneGUI()
```

It worth noting that most of the dependencies of the package are on CRAN. Note that in case of setting the repository with the setRepositories() command instead of the biocLite() function, we need to select both Bioconductor and CRAN in order to install the package properly.

The main window of the package is shown in Figure 10.2. The top tab lists several submenus. First the submenu 'File' (A in Figure 10.2) allows to load the data set and to display the data values as table. The data compatible with package can be provided either as plain text file, Microsoft Excel spreadsheet or the .RData file. The submenu 'Analysis' (B) comprises the methods for inference, estimation and model selection, i.e. it contains the packages IsoGene, orQA and goric. The clustering of the genes based on their profiles can be performed in a separate submenu (C), using the methods implemented in ORCME and ORIClust. Some of the plots can be obtained from the analysis windows, but more general plots are listed in the visualization techniques submenu (D). The graphical techniques listed in submenu D typically use outputs of the methods implemented in other submenus. The plots can be saved in multiple file types. The last submenu 'Help' (E) contains the help files for the IsoGene package, the IsoGeneGUI package and the vignette for IsoGeneGUI. The box in the center of the main window (F) gathers the results of the analyses and displays summary statistics of the results. Additionally, it serves as indicator of which outputs are currently active (if analysis was run multiple times) and will be plotted by visualization tools.

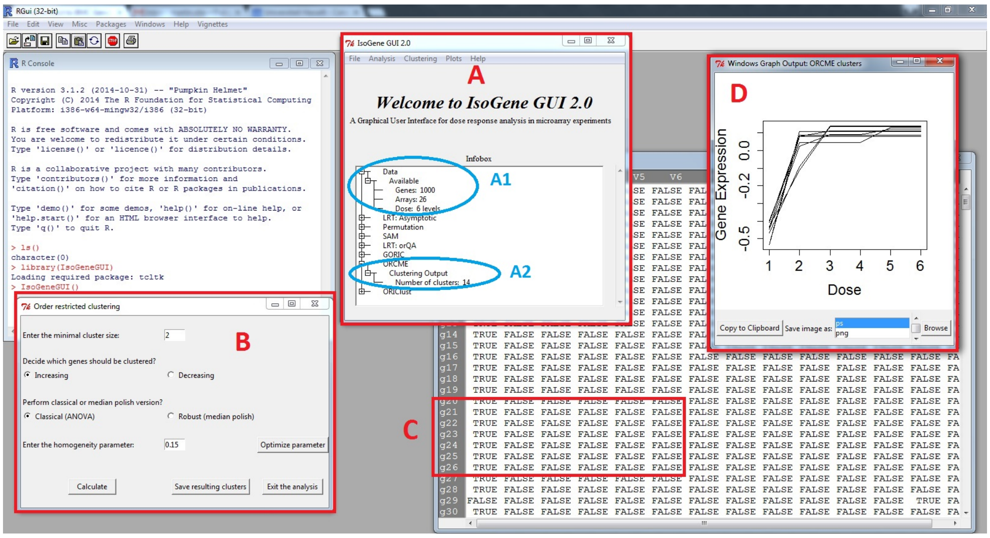
An example of the package interface is fully shown in Figure 10.3. We can see the main window again (A), now with the box showing the properties of active data set (A1) and a summary of results of a clustering procedure (A2). The window that was used for clustering with ORCME method is displayed on the left side of the Figure 10.3 (B) and the results are displayed in the table (C). One of the clusters was plotted using one of the visualization options (D). Further examples are shown in following section.



Welcome to IsoGene GUI 2.0

A Graphical User Interface for dose response analysis in microarray experiments

Figure 10.2: The `IsoGeneGUI` package main menu with highlighted submenus.



```
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> ls()
character(0)
> library(IsoGeneGUI)
Loading required package: tcltk
> IsoGeneGUI()
```

Welcome to IsoGene GUI 2.0

A Graphical User Interface for dose response analysis in microarray experiments

Data
  Available
    Genes: 1000
    Arrays: 26
    Dose: 6 levels
LRT: Asymptotic
Permutation
SAM
LRT: orQA
GORIC
ORCME
  Clustering Output
    Number of clusters: 14
ORIClust

Enter the minimal cluster size: 2

Decide which genes should be clustered?
● Increasing ○ Decreasing

Perform classical or median polish version?
● Classical (ANOVA) ○ Robust (median polish)

Enter the homogeneity parameter: 0.15

```
V5    V6
g14  TRUE FALSE FALSE FALSE FALSE FALSE FAL
g15  TRUE FALSE FALSE FALSE FALSE FALSE FAL
g16  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g17  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g18  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g19  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g20  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g21  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g22  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g23  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g24  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g25  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g26  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g27  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g28  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
g29 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FA
g30  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
```

Figure 10.3: `R` with opened `IsoGeneGUI` package.

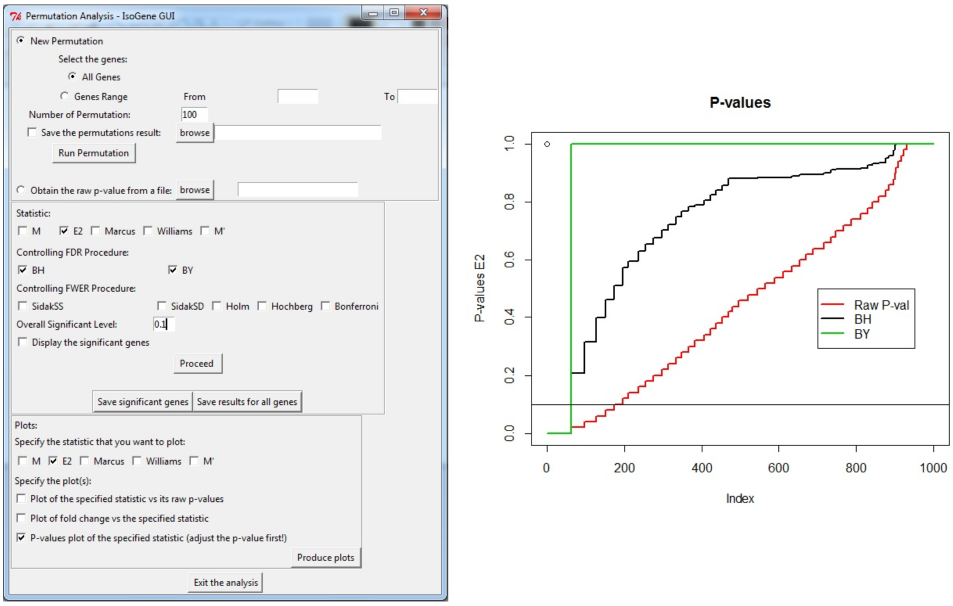
10.5 Applications

The IsoGeneGUI implementation of the available methods is less flexible than in original packages. That is natural trade-off between clarity and accessibility of options in GUI compared to plain R packages that are more flexible but also more difficult to operate without proficient experience with R. This section describes the implementation of the methods for inference, clustering and model selection. The examples shown in Figure 10.4 to Figure 10.7 were obtained using the example data set dopamine that is part of the IsoGeneGUI package. In each figure, one method is presented, accompanied with one of available graphical displays.

10.5.1 Inference

The permutation test is implemented for all five test statistics discussed above, using the functions from IsoGene package. For the LRT, a much faster implementation of the permutation test is available from orQA. Both methods produce the same result (within the sampling error), so the slower version should be used only in case that additional test statistics are of interest. Additionally, there is an asymptotic solution available for the LRT as well. Note that it is advised to avoid this option in case of small sample sizes.

The window that facilitates permutation test based on the IsoGene package is shown in Figure 10.4. The left panel shows the window itself. The top part allows to select the genes for which the raw p-values based on permutation test will be obtained. The middle part of window offers seven multiplicity adjustment methods and computation of significant genes based on any of the five test statistics. The last part produces three types of plots. The right panel of Figure 10.4 shows an example of one of the plots: the adjustment of p-value while controlling FDR. In this case, both BH and BY methods agreed on same set of genes, but that is not necessarily case in general. For FDR equal to 5%, we expect three false discoveries among the 62 null hypotheses that were rejected. The left panel of Figure 10.5 shows the window for the LRT using the orQA package, providing nearly same options as permutation method. The right panel of Figure 10.5 shows example of so called 'volcano plot' that compares the -log(p-value) and fold change. Note that the high value for -log(p-value) of genes with fold change around zero is often caused by a small variance among the observations of these genes. This is an indication that the SAM method should be applied (Lin *et al.*, 2012d).



Permutation Analysis - IsoGene GUI

● New Permutation

Select the genes:

● All Genes

○ Genes Range From To

Number of Permutation: 100

□ Save the permutations result: browse

Run Permutation

○ Obtain the raw p-value from a file: browse

Statistic:

□ M ☑ E2 □ Marcus □ Williams □ M'

Controlling FDR Procedure:

☑ BH ☑ BY

Controlling FWER Procedure:

□ SidakSS □ SidakSD □ Holm □ Hochberg □ Bonferroni

Overall Significant Level: 0.1

□ Display the significant genes

Proceed

Save significant genes | Save results for all genes

Plots:

Specify the statistic that you want to plot:

□ M ☑ E2 □ Marcus □ Williams □ M'

Specify the plot(s):

□ Plot of the specified statistic vs its raw p-values

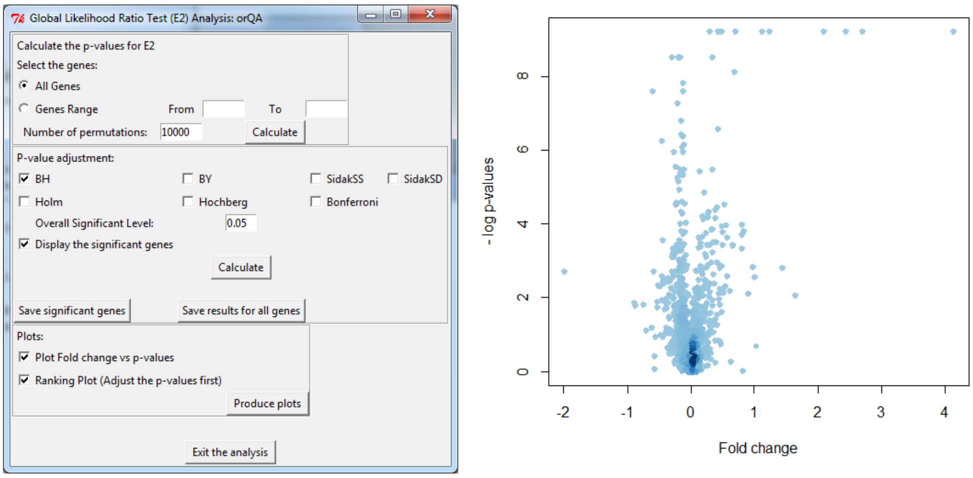
□ Plot of fold change vs the specified statistic

☑ P-values plot of the specified statistic (adjust the p-value first!)

Produce plots

Exit the analysis

Figure 10.4: Resampling based inference. Left panel: The window for performing permutation test. Right panel: Plot of an effect of multiplicity adjustment.



Global Likelihood Ratio Test (E2) Analysis: orQA

Calculate the p-values for E2

Select the genes:

● All Genes

○ Genes Range From To

Number of permutations: 10000 Calculate

P-value adjustment:

☑ BH □ BY □ SidakSS □ SidakSD

□ Holm □ Hochberg □ Bonferroni

Overall Significant Level: 0.05

☑ Display the significant genes

Calculate

Save significant genes | Save results for all genes

Plots:

☑ Plot Fold change vs p-values

☑ Ranking Plot (Adjust the p-values first)

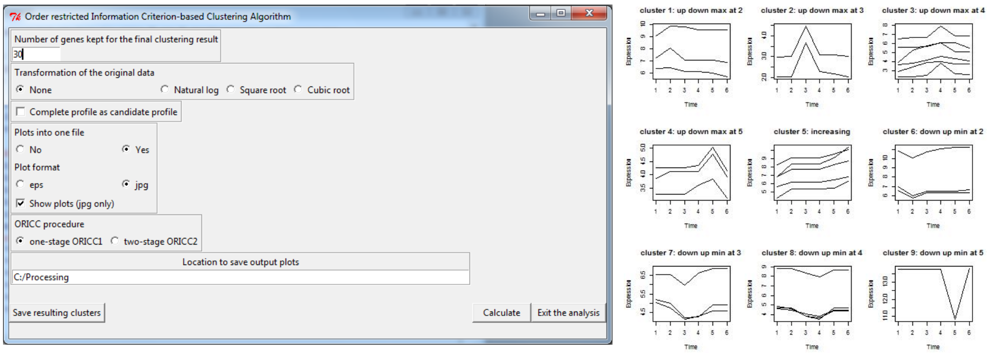
Produce plots

Exit the analysis

Figure 10.5: Inference with `orQA`. Left panel: The window for performing LRT. Right panel: Volcano plot.

10.5.2 Clustering

Order restricted clustering is addressed by two algorithms, the δ-clustering from `ORCME` and the ORICC from `ORIClust`. As mentioned above, the package contains two versions of the δ-clustering method: clustering based on the least squares and a robust clustering based on least of absolute residuals. The ORCME window and output is shown in Figure 10.3. The window implementing ORICC is shown in left panel of Figure 10.6. All monotone and umbrella profiles are automatically considered and the user cannot influence this setting. However, this setting provides the flexible framework for clustering. The complete profile can be included to the set as well. One or two-stage type of ORICC can be run and output is automatically saved in both text and visual form. The clustering results are shown in right panel of Figure 10.6 for case in which the top 30 genes are kept for final clustering step.



Order restricted Information Criterion-based Clustering Algorithm

Number of genes kept for the final clustering result: 30

Transformation of the original data: ● None ○ Natural log ○ Square root ○ Cubic root

☐ Complete profile as candidate profile

Plots into one file: ○ No ● Yes

Plot format: ○ eps ● jpg

☑ Show plots (jpg only)

ORICC procedure: ● one-stage ORICC1 ○ two-stage ORICC2

Location to save output plots: C:/Processing

Save resulting clusters
Calculate
Exit the analysis
cluster 1: up down max at 2
cluster 2: up down max at 3
cluster 3: up down max at 4
cluster 4: up down max at 5
cluster 5: increasing
cluster 6: down up min at 2
cluster 7: down up min at 3
cluster 8: down up min at 4
cluster 9: down up min at 5
Expression
Time

Figure 10.6: Order restricted clustering using `ORIClust`. Left panel: The window for clustering. Right panel: Plot of all the resulting clusters.

10.5.3 Model selection

The current implementation in `IsoGeneGUI` runs automatically GORIC for all possible models for a given direction (upward or downward trends). Therefore, for an experiment with control and $K-1$ dose levels, 2^{K-1} models are considered, including the null model of no dose effect. In case that some of these models are not considered for the analysis, the posterior weights can be easily normalized for the smaller set of models. Only one gene at the time can be analyzed using the GORIC procedure, due to computational intensity of the derivation of the model weights. The GORIC window is shown in left panel of Figure 10.7. For the `dopamine` data, there are six dose levels and therefore, for an upward trend that are 32 possible monotone non-decreasing models (including the null model). We focus on the results obtained for gene *156_at* (row 56 in the data set).

The middle plot of Figure 10.7 displays the data and the model with highest weights, M_{15}, increasing in all doses except the last one. The right panel shows the weights for all models, revealing that there are two models with almost equal weights (M_{15} and M_{31}). The difference between them is the the equality or increment between last two doses.

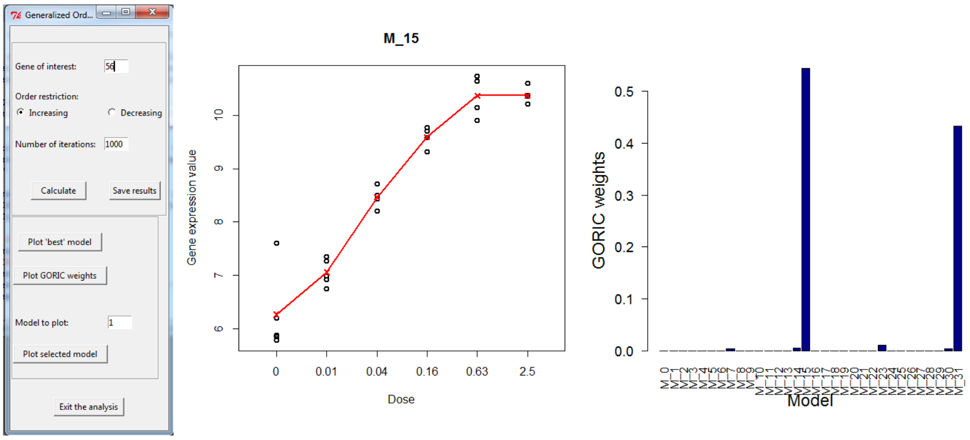

Gene of interest: 56
Order restriction: ● Increasing ○ Decreasing
Number of iterations: 1000
Model to plot: 1

Figure 10.7: The GORIC method. Left panel: The window for performing analysis for one gene. Middle panel: Dose-response relationship under model M_{15} for gene *156_at*. Right panel: GORIC weights for all the models fitted to gene *156_at*.

10.6 Summary

The analysis of dose-response relationship for order restricted experiments is highly relevant in the drug discovery process. Multiple R packages offer methodology within this framework. The new version of the IsoGeneGUI package encompasses a wide range of these packages in an unified way. The package contains data analysis tools for estimation, inference, model selection and clustering. To our knowledge, it is the only software package providing such a wide range of tools simultaneously. Additionally, the GUI implementation of the package allows non-statisticians to conduct the analysis with only minimal knowledge of R. In summary, the package IsoGeneGUI is a state-of-the-art collection of methodologies covering a wide range of analyses that are meaningful for order restricted microarray experiments. Moreover, the package can be used in a straightforward way by the general scientific community.

Chapter 11

Discussion

The dose-response modelling is a common theme of this thesis, with a special focus on the monotonicity assumption. Various aspects of dose-response modelling were introduced, including estimation, model selection, inference, model complexity, clustering and model uncertainty. We focused on methods suitable to be applied in microarray experiments that are typical representatives of high dimensional problems. Multiple approaches to analysis were explored and evaluated, grounded either in frequentist or Bayesian statistics frameworks. The research of new methodologies focused mainly on Bayesian methods, while frequentist procedures were used in applied work and software development.

Most of the methods introduced in the thesis are developed under order constraints. They benefit from this partial knowledge about the dose-response shape and, as mentioned in introduction of the thesis, there is wide range of applications when such an assumption is suitable. However, most of the methodology presented here can be extended beyond the order constraints, as in Chapter 8 where fractional polynomials are used. The value of the presented analysis framework is the generality and flexibility. Straightforward modifications allow for application in varying settings, without need of a tedious theoretical development. This is especially evident in case of BVS, where changes in distributional assumptions or prior knowledge are conducted via changes in hyperparameters' distributions. Analogously, analysis framework presented in second part of the thesis can be modified by replacing statistical tests, filtering methods or integration methods to adjust for particular area of application.

11.1 Bayesian variable selection

The first part of the thesis was focused on Bayesian variable selection modelling and demonstrated its suitability for conducting inference, estimation and model selection. Our aim was to develop a complete Bayesian framework for order restricted analysis of dose-response experiments which will be comparable to the LRT and the MCTs that are commonly used in this setting. We consider the BVS concept appealing for two main reasons. Firstly, it is possible to incorporate any available prior information in the model. A better utilization of information obtained in the previous experiments is important in current scientific reasoning. Data storage and linkage of different data sources are currently developing rapidly, facilitating the use of prior information in daily practice. For example, in drug development process, the data from early discovery stages can be incorporated in later stages of the development. The second reason is the unified data analysis framework provided by the BVS. The posterior model probabilities are obtained simultaneously with the estimates for the unknown parameters of the model. When model specific estimates are of interest, they can be obtained from the MCMC as transformed variables. The overall performance of the BVS was shown to be as good as any competing method in terms of inference and better in terms of model selection for specific profiles.

In the research presented in this thesis, the BVS was applied to independent and normally distributed data. The BVS can be extended to more complex setting, i.e. incorporate multiple responses and their interrelationship, to take into account random effects for correlated data or to use of non-Gaussian distributions. Applied in a microarray setting, the posterior model probabilities can be used to cluster genes, rather than to relay on the best model only. The computational speed for the basic BVS model is favourable as well. The computational time will increase significantly, if permutation test is applied. Nevertheless, the problem is 'embarrassingly parallel', i.e. suitable to be run on multiple cores. In fact, most of the simulations introduced in the thesis were conducted using supercomputers. However, it may remain challenging how to deliver the method to a broad scientific public that does not have access to such resources. Ongoing research suggests that the computational time will cease to be an issue in close future and the BVS model may be applied on daily basis on high dimensional data sets. Alternative option is the use of cloud computing. The conversion of the BVS model into user friendly environment in R that would be suitable for application in high dimensional setting is currently an ongoing research line.

The BVS framework provides interesting challenges for future research. Firstly, the specification of the variable selection component of the model can be modified. The approach introduced in this thesis followed Kuo and Mallick (1998), with independent variables δ_h and z_h. The advantage of this approach is absence of any tuning parameter, but it suffers from poor mixing of MCMC chains if the prior on δ_h is too vague. In our case, truncation of the non-informative prior distribution of δ_h with constant A solves this issue, but implies necessity of the determination of constant A in a way that does not influence the results of the analysis. There is often enough knowledge about range of possible values of response to set A prior to experiment takes place. Alternatively, A can be set as the maximum of response in the data set or maximal increment observed in the data.

Alternative specification of priors introduce dependency between δ_h and z_h. Stochastic search variable selection (George and McCulloch, 1993) can be applied that specifies δ_h as following a mixture of normal distributions, i.e. $\delta_h \sim (1 - z_h)N(0, \tau^2) + z_h N(0, \kappa\tau^2)$. Truncation in zero will be added in our case. Parameters τ^2 and κ are data dependent and need to be tuned by the user. Such tuning may be complicated task and it is not applicable for high dimensional data, but it can be applied for the single experiment setting. Elaborated discussion about possible specification of variable selection can be found in O'Hara and Sillanpää (2009).

The prior independence among increments δ_h, $h = 1, \ldots, K$ can be relaxed, as seen for example in Ohlssen and Racine (2015). The overall increment δ can be sampled first and then separated into the dose-specific increments. Naturally, this approach reflects the reality better than independence of increments, but its disputable if there is any practical gain based on this change. Additionally, the prior distributions of hyperparameters can be modified in order to improve the analysis. The hyperparameters for distribution of δ_h may be omitted and non-informative distribution for δ_h themselves can be used. Alternatively, 'weakly informative priors' introduced by Gelman (2006) can be used for hyperparameters. Such priors use half-t family of distributions instead of inverse-gamma distribution. They imply very vague information about the parameter, but simultaneously allow to restrict for feasible parameter values only. Finally, if a strong prior scientific knowledge is available, investigation of most efficient way how to incorporate it in the BVS model can be an interesting topic for a future research.

Another topic for future development is related to posterior expected complexity pEC. A model selection procedure within the Bayesian framework is a challenging task. The DIC is commonly used, but it suffers from several issues, as dependency on reparametrization, lack of consistency and generally weak theoretical justification of the criterion (Spiegelhalter *et al.*, 2014). Hence, its suitability is limited. The BVS model does not suffer from such problems, because the variable selection approach uses indicator variables z_h that unambiguously determine particular model. Therefore, posterior model probabilities can be obtained and consequently used for model selection. However, such an approach works only if a one-to-one relationship holds between configurations of the z and the candidate models. This is not a case if two different sets of models should be compared or if the set of models is not hierarchically structured. In former case, the posterior expected complexity may lead to a solution, representing complexity given the set of models, the data and prior distribution, while accounting for model uncertainty (within each set of models). The posterior complexity for each set may be computed, transformed into the information criterion by adding the likelihood term and compared in order to select which set of models should be preferred. However, the dependency of such criterion on prior distribution needs to be carefully investigated. Additionally, it is not clear how to use pEC to compare individual models that are not hierarchically structured.

As we mentioned before, the main goal in the first part of the thesis was to develop the BVS model as an alternative to frequentist analysis of dose-response experiment, while benefiting from Bayesian framework characteristics. In terms of inference, there is a connection between multiple contrast tests and the BVS model. While achieving generally high power, the BVS provides complete information about posterior distribution of the models being the true underlying model. The flexibility of BVS allows to extend the framework to topics that are typically addressed by frequentist method. For example, the ratio parameters can be modelled instead of dose-specific means, as in Lin *et al.* (2012a). The ratio of dose-specific means is important quantity because of its biological interpretation as relative effect of the dose compared to the baseline value. Biological significance may be incorporated in the analysis by testing if the ratio of the means is higher (or lower) than some prespecified value $\omega > 1$ ($\omega < 1$, respectively).

As mentioned above, clustering of genes in microarray experiments based on information criteria (Lin *et al.*, 2012c) can be addressed via posterior model probabilities. In its simplest form, gene can be grouped according to the model that has highest posterior probability. Additionally, we can cluster genes based on whole distribution of posterior model probabilities. For each gene, there is vector of posterior model probabilities than can be used in order to cluster together genes with similar distribution of probability of being true model across all possible models. This clustering is inherently different than clustering based on best model only and also than clustering based on dose-specific means that does not take into account variability. Alternatively, biclustering methods (Madeira and Oliviera, 2004 and Kasim *et al.*, to be published 2016) can be applied. The resulting biclusters contain genes that have same posterior model probability for subset of models, up to a multiplicative constant. Biclustering is useful if only similarity over subset of models is important and in case that only the ratios of certain model probabilities are of interest.

Finally, the method for specification of the threshold using the conditional FDR (Newton *et al.*, 2007) was described in Section 2.5.3. As discussed in Section 2.7, the cFDR control does not extend to the FDR control, so the interpretation of the results of the cFDR method interpretation can be confusing. The method that would determine the threshold based on the FDR control would be very useful tool in a high dimensional setting. In this thesis, the permutation test was used instead which allows for the FDR control using the resulting p-values. The method that would allow for FDR control using $P(g_r|\text{data})$ itself could lead to significant reduction of computational speed.

11.2 Toxicogenomics

The analysis of Japanese Toxicogenomics Project data sets provides an analysis workflow for translational research. It can be extended to any platform and context and different modules in the framework can be modified (e.g. underlying model, multiplicity correction, performed statistical tests, selection procedure, clustering methods, etc.). The main aim of the second part of the thesis was to illustrate such complete framework including the interpretation of the results. The framework is of exploratory nature, so the results need further evaluation, using scientific knowledge and follow-up experiments. If group of translatable/disconnected genes would be identified with high confidence, a further extension of the workflow could be the prediction of the effects on one of the platforms/species using the other. Such a step could reduce the number of animals per experiment and/or the number of experiments.

11.3 Software development

A development of a methodology for data analysis needs to be accompanied by providing of the acquired knowledge to the scientific community, both in terms of publication and software products. Therefore, multiple software packages were produced in order to support the development, with main representatives being `ORCME` package and `IsoGeneGUI` package. The `ORCME` package provides cluster analysis based on the δ-clustering method. It is flexible tool for exploratory analysis of microarray data, identifying genes with similar profiles.

The development of the second generation of the `IsoGeneGUI` package was based on a community based software development. The idea was to develop an envelope package that includes all available software in `R` related to the analysis of dose-response experiments. The simplicity of `GUI` allows it to be used by researches with limited knowledge of `R` and therefore to spread the valuable methodology beyond the borders of statistical community. The envelope nature ensures that state-of-the-art methodology is available, putting together various groups of researches and offering as complete tool as possible.

Both concepts can be extended further. The GUI can be integrated into the ready-to-use platforms, such as RCommander (Fox, 2005), that makes their use even simpler. Recent development of Shiny framework (Chang *et al.*, 2015) allows to create R based interactive applications that are run in web browser. Typically, they can reside on server, so distant user can access them without necessity of installing R or knowing how to work with it. Naturally, such approaches can reduce flexibility, but support user friendly tools suitable for scientists without deep knowledge of statistics. Envelope packages demonstrate collective effort of broader scientific community. Therefore, it may be challenging to maintain them as time goes by. The changes in packages that envelope package depends upon may imply modification of the envelope package. Similarly, post hoc addition of new package into the envelope package may be demanding. Solution to these issues is standardization of the development processes and communication of standard development procedures to the scientific community. Then, the preparation of the codes can be done by authors of particular packages instead of maintainer of the envelope packages who merely combines all materials. Excellent example of flexible standards for envelope packages can be seen in the REST R package (De Troyer, 2015).

Bibliography

Afshari, C. A., Hamadeh, H. and Bushel, P. R. (2011) The evolution of bioinformatics in toxicology: advancing toxicogenomics. *Toxicological Sciences*, **120**, S225–237.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Amaratunga, D., Cabrera, J. and Kovtun, V. (2008) Microarray learning with ABC. *Biostatistics*, **9**, 128–136.

Amaratunga, D., Cabrera, J. and Shkedy, Z. (2014) *Exploration and Analysis of DNA Microarray and Other High-Dimensional Data*. John Wiley & Sons, 2nd edn.

Ames, B. N., McCann, J. and Yamasaki, E. (1975) Methods for detecting carcinogens and mutagens with the salmonella/mammalian-microsome mutagenicity test. *Mutation Research*, **31**, 347–364.

Andrews, N. A. (2013) Skeletal regulation of glucose metabolism: challenges in translation from mouse to man. *IBMS BoneKEy*, **10**, 1.

Anraku, K. (1999) An information criterion for parameters under a simple order restriction. *Biometrika*, **86**, 141–152.

Arnodl, K., Gosling, J. and Holmes, D. (2000) *The Java Programming Language*. Prentice Hall, third edn.

Arrowsmith, J. (2011) Trial watch: phase III and submission failures: 2007-2010. *Nature Reviews Drug Discovery*, **10**, 87.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Bajorath, J. (2001) Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. *Drug Discovery Today*, **6**, 989–995.

Barlow, R. E., Bartholomew, D. J., Bremner, M. J. and Brunk, H. D. (1972) *Statistical Inference under Order Restriction*. New York: John Wiley & Sons.

Bartholomew, D. (1961) Ordered tests in the analysis of variance. *Biometrika*, **48**, 325–332.

Baum, P., Schmid, R., Ittrich, C., Rust, W., Fundel-Clemens, K., Siewert, S., Baur, M., Mara, L., Gruenbaum, L., Heckel, A., Eils, R., Kontermann, R. E., Roth, G. J., Gantner, F., Schnapp, A., Park, J. E., Weith, A., Quast, K. and Mennerich, D. (2010) Phenocopy-a strategy to qualify chemical compounds during hit-to-lead and/or lead optimization. *PloS One*, **5**, e14272.

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

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165–1188.

Bijnens, L., Göhlmann, H. W. H., Lin, D., Talloen, W., Perera, T., Van Den Wyngaert, I., De Ridder, F., De Bondt, A. and Peeters, P. (2012) Functional genomics dose-response experiments. In: *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R* (Eds. D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga and L. Bijnens), 69–80. Springer.

Bonferroni, C. E. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Bornkamp, B., Pinheiro, J. C. and Bretz, F. (2009) MCPMod - an R package for the design and analysis of dose-finding studies. *Journal of Statistical Software*, **29**, 1–23.

Breiman, L. (1996) Random forests. *Machine Learning*, **24**, 123–140.

Bretz, F. (1999) *Powerful Modification on Williams' Test on Trend*. Ph.D. thesis, Universität Hannover.

Bretz, F. (2006) An extension of the Williams trend test to general unbalanced linear models. *Computational Statistics and Data Analysis*, **50**, 1735–1748.

Bretz, F. and Hothorn, L. A. (2003) Statistical analysis of monotone or non-monotone dose-response data from *in vitro* toxicological assays. *Alternatives to Lab Animals*, **31**, 81–96.

Bretz, F., Pinheiro, J. C. and Branson, M. (2005) Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, **61**, 738–748.

Briggs, K., Cases, M., Heard, D. J., Pastor, M., Pognan, F., Sanz, F., Schwab, C. H., Steger-Hartmann, T., Sutter, A., Watson, D. K. and Wichard, J. D. (2012) Inroads to predict *in vivo* toxicology - an introduction to the eTOX project. *International Journal of Molecular Sciences*, **13**, 3820–3846.

Brune, K. and Patrignani, P. (2015) New insights into the use of currently available non-steroidal anti-inflammatory drugs. *Journal of Pain Research*, **8**, 105–118.

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997) Model selection: An integral part of inference. *Biometrics*, **53**, 603–618.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information - Theoretic Approach*. New York: Springer.

Burnham, K. P. and Anderson, D. R. (2004) Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33**, 261–304.

Calinski, R. B. and Harabasz, J. (1974) A dendrite method for cluster analysis. *Communications in Statistics*, **3**, 1–27.

Casella, G. and Moreno, E. (2006) Objective Bayesian variable selection. *Journal of the American Statistical Association*, **101**, 157–167.

Castellsague, J., Riera-Guardia, N., Calingaert, B., Varas-Lorenzo, C., Fourrier-Reglat, A., Nicotra, F., Sturkenboom, M. and Perez-Gutthan, S. (2012) Individual NSAIDs and upper gastrointestinal complications: A systematic review and meta-analysis of observational studies (the SOS project). *Drug Safety*, **35**, 1127–1146.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y. and McPherson, J. (2015) *shiny: Web Application Framework for R*. URL `http://CRAN.R-project.org/package=shiny`. R package version 0.11.1.

Chen, M., Vijay, V., Shi, Q., Liu, Z., Fang, H. and Tong, W. (2011) FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discovery Today*, **16**, 697–703.

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

Claeskens, G. and Hjort, N. L. (2008) *Model Selection and Model Averaging*. Cambridge University Press.

Clevert, D.-A., Heusel, M., Mitterecker, A., Talloen, W., Göhlmann, H. W. H., Wegner, J., Mayr, A., Klambauer, G. and Hochreiter, S. (2012) Exploiting the Japanese Toxicogenomics Project for predictive modelling of drug toxicity. In: *CAMDA 2012, Satellite Meeting of ISMB/ECCB 2012, Long Beach CA, USA, July 13-14*.

Clyde, M. and George, E. I. (2004) Model uncertainty. *Statistical Science*, **19**, 81–94.

Davidov, E., Holland, J., Marple, E. and Naylor, S. (2003) Advancing drug discovery through systems biology. *Drug Discovery Today*, **8**, 175–183.

De Troyer, E. (2015) *REST: RcmdrPlugin Easy Script Templates*. URL `http://CRAN.R-project.org/package=REST`. R package version 1.0.1.

Dellaportas, P., Forster, J. J. and Ntzouras, I. (2002) On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**, 27–36.

Denwood, M. J. (In Review) runjags: An R package providing interface utilities, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*. URL `http://cran.r-project.org/web/packages/runjags/`.

Djira, G. D., Hasler, M., Gerhard, D. and Schaarschmidt, F. (2012) *mratios: Inferences for ratios of coefficients in the general linear model*. URL `http://CRAN.R-project.org/package=mratios`. R package version 1.3.17.

Do, K. A., Müller, P. and Vannucci, M. (2006) *Bayesian inference for gene expression and proteomics*. Cambridge University press, London , UK.

Dunnett, C. W. (1955) A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50**, 1096–1121.

Dunson, D. B. and Neelon, B. (2003) Bayesian inference on order constrained parameters in generalized linear models. *Biometrics*, **59**, 286–295.

Dwass, M. (1957) Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, **28**, 181–187.

Ellinger-Ziegelbauer, H., Gmuender, H., Bandenburg, A. and Ahr, H. J. (2008) Prediction of a carcinogenic potential of rat hepatocarcinogens using toxicogenomics analysis of short-term *in vivo* studies. *Mutation Research*, **637**, 23–39.

Emerson, J. D. and Hoaglin, D. C. (1983) Analysis of two-way tables by medians. In: *Understanding Robust and Exploratory Data Analysis* (Eds. D. C. Hoaglin, F. Mosteller and J. W. Tukey), 165–210. John Wiley & Sons.

Enayetallah, A. E., Puppala, D., Ziemek, D., Fischer, J. E., Kantesaria, S. and Pletcher, M. T. (2013) Assessing the translatability of *in vivo* cardiotoxicity mechanisms to *in vitro* models using causal reasoning. *BMC Pharmacology and Toxicology*, **14**, 1–12.

Eren, K., Deveci, M., Küçüktunç, O. and Çatalyürek, U. V. (2013) A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*, **14**, 279–292.

Ernst, J. and Bar-Joseph, Z. (2006) STEM: A tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, **7**, 191.

Ernst, M. D. (2004) Permutation methods: A basis for exact inference. *Statistical Science*, **19**, 676–685.

European Medicines Agency (2002) *Points to Consider on Multiplicity Issues in Clinical Trials*, vol. CPMP/EWP/908/99. London: Committee for Proprietary Medicinal Products.

Fanton, C. P., Rowe, M. W., Moler, E. J., Ison-Dugenny, M., De Long, S. K., Rendahl, K., Shao, Y., Slabiak, T., Gesner, T. G. and MacKichan, M. L. (2006) Development of a screening assay for surrogate markers of CHK1 inhibitor-induced cell cycle release. *Journal of Biomolecular screening*, **11**, 792–806.

Fisher, R. A. (1936) "The coefficient of racial likeness" and the future of craniometry. *Journal of Royal Anthropological Institute of Great Britain and Ireland*, **66**, 57–63.

Fitzhugh, O. G., Nelson, A. A. and Quaife, M. L. (1964) Chronic oral toxicity of aldrin and dieldrin in rats and dogs. *Food Cosmetic Toxicology*, **2**, 551–562.

Food and Drug Administration (2004) Innovation or stagnation? Challenge and opportunity on the critical path to new medicinal products. U.S. Department of Health and Human Services.

Fox, J. (2005) The r commander: A basic-statistics graphical user interface to r. *Journal of Statistical*, **14**, 1–42.

Ganter, B., Tugendreich, S., Pearson, C. I., Ayanoglu, E., Baumhueter, S., Bostian, K. A., Brady, L., Browne, L. J., Calvin, J. T., Day, G. J., Breckenridge, N., Dunlea, S., Eynon, B. P., Furness, L. M., Ferng, J., Fielden, M. R., Fujimoto, S. Y., Gong, L., Hu, C., Idury, R., Judo, M. S., Kolaja, K. L., Lee, M. D., McSorley, C., Minor, J. M., Nair, R. V., Natsoulis, G., Nguyen, P., Nicholson, S. M., Pham, H., Roter, A. H., Sun, D., Tan, S., Thode, S., Tolley, A. M., Vladimirova, A., Yang, J., Zhou, Z. and Jarnagin, K. (2005) Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *Journal of Biotechnology*, **119**, 219–244.

Ge, Y., Dudoit, S. and Speed, T. P. (2003) Resampling-based multiple testing for microarray data analysis. *Test*, **12**, 1–77.

Gelfand, A. E., Smith, A. F. M. and Lee, T. M. (1992) Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, **87**, 523–532.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*. CRC Press.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Gentleman, R. C., Carey, V. J., Bates, D. M. and others (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

Gerhard, D. and Kuiper, R. M. (2012) *goric: Generalized Order-Restricted Information Criterion*. R package version 0.0-7.

Goldstein, M. (2006) Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, **1**, 403–420.

Hartigan, J. A. and Wong, M. A. (1979) Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **28**, 100–108.

Heijne, W. H., Jonker, D., Stierum, R. H., van Ommen, B. and Groten, J. P. (2005) Toxicogenomic analysis of gene expression changes in rat liver after a 28-day oral benzene exposure. *Toxicogenomics in Genetic Toxicology and Hazard Determination*, **575**, 85–101.

Hobin, J. A., M., D. A., Bockman, R., Cohen, S., Dechow, P., Eng, C., Galey, W., Morris, M., Prabhakar, S., Raj, U., Rubenstein, P., Smith, J. A., Stover, P., Sung, N., Talman, W. and R., G. (2012) Engaging basic scientists in translational research: identifying opportunities, overcoming obstacles. *Journal of Translational Medicine*, **10**, 72.

Hochberg, Y. and Benjamini, Y. (1990) More powerful procedures for multiple significance testing. *Statistics in Medicine*, **9**, 811–818.

Hochberg, Y. and Tamhane, A. C. (1987) *Multiple comparison procedures*. New York: Wiley.

Hochreiter, S., Clevert, D.-A. and Obermayer, K. (2006) A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**, 943–949.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Hornik, K. (2012) The comprehensive R archive network. *WIREs Computational Statistics*, **4**, 394–398.

Hothorn, L. A. (2006) Multiple comparisons and multiple contrasts in randomized dose-response trials-confidence interval oriented approaches. *Journal of Biopharmaceutical Statistics*, **16**, 711–731.

Hothorn, L. A. and Hauschke, D. (2000) Identifying the maximum safe dose: A multiple testing approach. *Journal of Biopharmaceutical Statistics*, **10**, 15–30.

Hothorn, T., Bretz, F. and Westfall, P. (2008) Simultaneous inference in general parametric models. *Biometrical Journal*, **50**, 346–363.

Hu, J., Kapoor, M., Zhang, W., Hamilton, S. and Coombes, K. (2005) Analysis of dose response effects on gene expression data with comparison of two microarray platforms. *Bioinformatics*, **21**, 3524–3529.

Jeffreys, H. (1961) *Theory of Probability*. London: Oxford University Press, 3rd edn.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**, 27–30.

Kasim, A., Otava, M. and Verbeke, T. (2014) *ORCME: Order Restricted Clustering for Microarray Experiments*. URL `http://CRAN.R-project.org/package=ORCME`. R package version 2.0.1.

Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S. and Talloen, W. (Eds.) (to be published 2016) *Applied Biclustering Methods for Big and High Dimensional Data Using R*. Chapman and Hall / CRC.

Kasim, A., Shkedy, Z. and Kato, B. S. (2012) Estimation and inference under simple order restrictions: hierarchical Bayesian approach. In: *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R* (Eds. D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga and L. Bijnens), 193–214. Springer, Berlin.

Kasim, A .and Lin, D., Van Sanden, S., Clevert, D.-A., Bijnens, L., Göhlmann, H. W. H., Amaratunga, D., Hochreiter, S., Shkedy, Z. and Talloen, W. (2010) Informative or non-informative calls for gene expression: a latent variable approach. *Statistical Applications in Genetics and Molecular Biology*, **9**, Article 4.

Kass, R. E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Kim, S. B., Kodell, R. L. and Moon, H. (2014) A diversity index for model space selection in the estimation of benchmark and infectious doses via model averaging. *Risk Analysis*, **34**, 453–464.

Kiyosawa, N., Manabe, S., Sanbuissho, A. and Yamoto, T. (2010) Gene set-level network analysis using a toxicogenomics database. *Genomics*, **96**, 39–49.

Klinglmueller, F. (2010) *orQA: Order Restricted Assessment Of Microarray Titration Experiments*. URL `http://CRAN.R-project.org/package=orQA`. R package version 0.2.1.

Klinglmueller, F., Tuechler, T. and Posch, M. (2011) Cross-platform comparison of microarray data using order restricted inference. *Bioinformatics*, **27**, 953–960.

Klugkist, I. and Hoijtink, H. (2007) The Bayes factor for inequality and about equality constrained models. *Computational Statistics and Data Analysis*, **51**, 6367–6379.

Klugkist, I. and Mulder, J. (2008) Bayesian estimation for inequality constrained analysis of variance. In: *Bayesian Evaluation of Informative Hypotheses* (Eds. H. H., I. Klugkist and P. A. Boelen), 27–52. New York: Springer.

Kodell, R. L. (2009) Replace the NOAEL and LOAEL with the $BMDL_{01}$ and $BMDL_{10}$. *Environmental and Ecological Statistics*, **16**, 9–12.

Kong, M., Rai, S. N. and Bolli, R. (2014) Statistical methods for selecting maximum effective dose and evaluating treatment effect when dose-response is monotonic. *Statistics in Biopharmaceutical Research*, **6**, 16–29.

Kuiper, R. M., Gerhard, D. and Hothorn, L. A. (2014) Identification of the minimum effective dose for normally distributed endpoints using a model selection approach. *Statistics in Biopharmaceutical Research*, **6**, 55–66.

Kuiper, R. M., Hoijtink, H. and Silvapulle, M. J. (2011) An Akaike-type information criterion for model selection under inequality constraints. *Biometrika*, **98**, 495–501.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86. URL `http://dx.doi.org/10.1214/aoms/1177729694`.

Kuo, L. and Mallick, B. (1998) Variable selection for regression models. *The Indian Journal of Statistics*, **60**, 65–81.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J. P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S. and Golub, T. R. (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes and disease. *Science*, **313**, 1929–1935.

Lin, D., Djira, G. D., Shkedy, Z., Burzykowski, T. and Hothorn, L. A. (2012a) Simultaneous inferences for ratio parameters using multiple contrasts test. In: *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R* (Eds. D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga and L. Bijnens), 233–247. Springer.

Lin, D., Hothorn, L. A., Djira, G. D. and Bretz, F. (2012b) Multiple contrast tests for testing dose-response relationships under order restricted alternatives. In: *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R* (Eds. D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga and L. Bijnens), 233–247. Springer.

Lin, D., Pramana, S., Verbeke, T. and Otava, M. (2013) *IsoGene: Testing for Monotonic Relationship between Gene Expression and Doses in a Microarray Experiment*. URL `http://CRAN.R-project.org/package=IsoGene`. R package version 1.0-22.

Lin, D., Shkedy, Z. and Aerts, M. (2012c) Classification of monotone gene profiles using information theory selection methods. In: *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R* (Eds. D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga and L. Bijnens), 151–164. Springer, Berlin.

Lin, D., Shkedy, Z., Burzykowski, T., Aerts, M., Göhlmann, H. W. H., De Bondt, A., Perera, T., Geerts, T., Van den Wyngaert, I. and Bijnens, L. (2009) Classification of trends in dose-response microarray experiments using information theory selection methods. *The Open Applied Informatics Journal*, **3**, 34–43.

Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D. and Bijnens, L. (Eds.) (2012d) *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R - Order Restricted Analysis of Microarray Data*. Springer-Verlag, Berlin. ISBN 978-3-642-24006-5.

Lin, D., Shkedy, Z., Yekutieli, D., Burzykowki, T., Göhlmann, H. W. H., De Bondt, A., Perera, T., Geerts, T. and Bijnens, L. (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**, Issue 1, Article 26.

Liu, J. (2010) Minimum effective dose. In: *Encyclopedia of Biopharmaceutical Statistics* (Ed. S. Chow), 799–800. Taylor & Francis, third edn.

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

Liu, T., Lin, N., Shi, N. and Zhang, B. (2012) *ORIClust: Order-restricted Information Criterion-based Clustering Algorithm*. URL `http://CRAN.R-project.org/package=ORIClust`. R package version 1.0-1.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

Madeira, S. C. and Oliviera, A. L. (2004) Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **1**, 24–45.

Marcus, R. (1976) The powers of some tests of the equality of normal means against an ordered alternative. *Biometrika*, **63**, 177–183.

McGonigle, P. and Ruggeri, B. (2014) Animal models of human disease: Challenges in enabling translation. *Biochemical Pharmacology*, **87**, 162–171.

Mestas, J. and Hughes, C. C. W. (2004) Of mice and not men: Differences between mouse and human immunology. *Journal of Immunology*, **172**, 2731–2738.

Miller, R. G. (1981) *Simultaneous Statistical Inference*. Springer.

Mosteller, F. and Tukey, J. W. (1977) *Data Analysis and Regression*. Addison-Wesley.

Mukerjee, H., Robertson, T. and Wright, F. T. (1987) Comparison of several treatments with a control using multiple contrasts. *Journal of the American Statistical Association*, **82**, 902–910.

Newton, M. A., Wang, P. and Kendziorski, C. (2007) Hierarchical mixture models for expression profiles. In: *Bayesian Inference for gene expression and proteomics* (Eds. K. M. Do, P. Müller and M. Vannucci), 40–52. Cambridge university press.

Neyman, J. and Pearson, E. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, **231**, 289–337.

Nie, A. Y., McMillian, M., Parker, J. B., Leone, A., Bryant, S., Yieh, L., Bittner, A., Nelson, J., Carmen, A., Wan, J. and Lord, P. G. (2006) Predictive toxicogenomics approaches reveal underlying molecular mechanisms of nongenotoxic carcinogenicity. *Molecular Carcinogenesis*, **45**, 914–933.

Ntzoufras, I. (2002) Gibbs variable selection using BUGS. *Journal of Statistical Software*, **7**, 1–19.

O'Hara, R. B. and Sillanpää, M. J. (2009) Review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, **4**, 85–118.

Ohlssen, D. and Racine, A. (2015) A flexible bayesian approach for modeling monotonic dose-response relationships in drug development trials. *Journal of Biopharmaceutical Statistics*, **25**, 137–156.

Otava, M., Shkedy, Z., Lin, D., Göhlmann, H. W. H., Bijnens, L., Talloen, W. and Kasim, A. (2014) Dose-response modeling under simple order restrictions using Bayesian variable selection methods. *Statistics in Biopharmaceutical Research*, **6**, 252–262.

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R. and Schacht, A. L. (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, **9**, 203–214.

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

Peddada, S. D., Lobenhofer, E., Li, L., Afshari, C., Weinberg, C. and D., U. (2003) Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics*, **19**, 834–841.

Pinheiro, J., Bornkamp, B., Glimm, E. and Bretz, F. (2014) Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine*, **33**, 1646–1661.

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

Plummer, M. (2003) Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Pognan, F. (2007) Toxicogenomics applied to predictive and exploratory toxicology for the safety assessment of new chemical entities: a long road with deep potholes. *Progress in Drug Research*, **64**, 219–238.

Pramana, S., Lin, D., Haldermans, P., Shkedy, Z., Verbeke, T., Göhlmann, H. W. H., De Bondt, A., Talloen, W. and Bijnens, L. (2010) Isogene: An R package for analyzing dose-response studies in microarray experiments. *The R Journal*, **2**.

Pramana, S., Lin, D., Haldermans, P. and Verbeke, T. (2012a) *IsoGeneGUI: A Graphical User Interface to Conduct a Dose-Response Analysis of Microarray Data*. URL `http://www.ibiostat.be/software/IsoGeneGUI/index.html`. R package version 1.20.0.

Pramana, S., Shkedy, Z., Göhlmann, H. W. H., Talloen, W., De Bondt, A., Straetemans, R., Lin, D. and Pinheiro, J. (2012b) Model-based approaches. In: *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R* (Eds. D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga and L. Bijnens), 215–232. Springer.

Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L. and Zitzler, E. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Raftery, A. E. (1995) Bayesian model selection in social research. *Sociological Methodology*, **25**, 111–163.

Ramsay, J. (1988) Monotone regression splines in action. *Statistical Science*, **3**, 425–441.

Rao, P. N. P. and Knaus, E. E. (2008) Evolution of nonsteroidal anti-inflammatory drugs (NSAIDs): Cyclooxygenase (COX) inhibition and beyond. *Journal of Pharmacy & Pharmaceutical Sciences*, **11**, 81s–110s.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988) *Order Restricted Statistical Inference.* John Wiley & Sons Ltd.

Rockova, V. and George, E. I. (2014) EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, **109**, 828–846.

Rockova, V., Lesaffre, E., Luime, J. and Löwenberg, B. (2012) Hierarchical Bayesian formulations for selecting variables in regression models. *Statistics in Medicine*, **31**, 1221–1237.

Rodríguez, L. A. G., Williams, R., Derby, L. E., Dean, A. D. and Jick, H. (1994) Acute liver injury associated with nonsteroidal anti-inflammatory drugs and the role of risk factors. *Archives of Internal Medicine*, **154**, 311–316.

Royston, P. and Altman, D. G. (1994) Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **43**, 429–467.

Rye, M. S., F., B. M., Cheeseman, M. T., Burgner, D., Blackwell, J. M., Brown, S. D. M. and Jamieson, S. E. (2011) Unraveling the genetics of otitis media: from mouse to human and back again. *Mammalian Genome*, **22**, 66–82.

Salton, G. (1988) *Automatic Text Processing*. Addison-Wesley.

Scannell, J. W., Blanckley, A., Boldon, H. and Warrington, B. (2012) Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, **11**, 191–200.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Seber, G. A. F. and Wild, C. J. (1989) *Nonlinear Regression*. New York: Wiley & Sons.

Seok, J., Warren, H. S., Cuenca, A. G., Mindrinos, M. N., Baker, H. V., Xu, W., Richards, D. R., McDonald-Smith, G. P., Gao, H., Hennessy, L., Finnerty, C. C., López, C. M., Honari, S., Moore, E. E., Minei, J. P., Cuschieri, J., Bankey, P. E., Johnson, J. L., Sperry, J., Nathens, A. B., Billiar, T. R., West, M. A., Jeschke, M. G., Klein, M. B., Gamelli, R. L., Gibran, N. S., Brownstein, B. H., Miller-Graziano, C., Calvano, S. E., Mason, P. H., Cobb, J. P., Rahme, L. G., Lowry, S. F., Maier, R. V., Moldawer, L. L., Herndon, D. N., Davis, R. W., Xiao, W., Tompkins, R. G. and the Inflammation and Host Response to Injury, Large Scale Collaborative Research Program (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences*, **110**, 3507–3512.

Shkedy, Z., Amaratunga, D. and Aerts, M. (2012a) Estimation under order restrictions. In: *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R* (Eds. D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga and L. Bijnens), 11–27. Springer, Berlin.

Shkedy, Z., Amaratunga, D. and Lin, D. (2012b) Testing of equality of means against ordered alternatives. In: *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R* (Eds. D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga and L. Bijnens), 29–42. Springer, Berlin.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2014) The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**, 485–493.

Straetemans, R. (2012) Nonlinear modeling of dose-response data. In: *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R* (Eds. D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga and L. Bijnens), 43–66. Springer.

Sugiura, N. (1978) Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, **7**, 13–26. URL `http://dx.doi.org/10.1080/03610927808827599`.

Suter, L., Schroeder, S., Meyer, K., Gautier, J. C., Amberg, A., Wendt, M., Gmuender, H., Mally, A., Boitier, E., Ellinger-Ziegelbauer, H., Matheis, K. and Pfannkuch, F. (2011)

EU framework 6 project: predictive toxicology (PredTox) - overview and outcome. *Toxicology and Applied Pharmacology*, **252**, 73–84.

Talloen, W., Clevert, D. A., Hochreiter, S., Amaratunga, D., Bijnens, L., Kass, S. and Göhlmann, H. W. H. (2007) I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, **23**, 2897–2902.

Talloen, W. and Göhlmann, H. W. H. (2009) *Gene Expression Studies Using Affymetrix Microarrays*. Chapman & Hall.

Tibshirani, R., Walther, G. and Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society B*, **63**, 411–423.

Timbrell, J. A., Delaney, J. and Waterfield, C. J. (1996) Correlation between *in vivo* and *in vitro* toxic effects of foreign compounds. *Comparative Haematology International*, **6**, 232–236.

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 United States of America*, **98**, 5116–5121.

Uehara, T., Kiyosawa, N., Shimizu, T., Omura, K., Hirode, M., Imazawa, T., Mizukawa, Y., Ono, A., Miyagishima, T., Nagao, T. and Urushidani, T. (2008) Species-specific differences in coumarin-induced hepatotoxicity as an example toxicogenomics-based approach to assessing risk of toxicity to humans. *Human & Experimental Toxicology*, **27**, 23–35.

Uehara, T., Ono, A., Maruyama, T., Kato, I., Yamada, H., Ohno, Y. and Urushidani, T. (2010) The Japanese Toxicogenomics Project: application of toxicogenomics. *Molecular Nutrition & Food Research*, **54**, 218–277.

Šidák, Z. (1971) On probabilities of rectangles in multivariate Student distributions: their dependence on correlations. *The Annals of Mathematical Statistics*, **42**, 169–175.

Wang, E.-J., Snyder, R. D., Fielden, M. R., Smith, R. J. and Gu, Y.-Z. (2008) Validation of putative genomic biomarkers of nephrotoxicity in rats. *Toxicology*, **246**, 91–100.

Wang, S.-J., Hung, H. M. J. and O'Neill, R. (2011) Regulatory perspectives on multiplicity in adaptive design clinical trials throughout a drug development program. *Journal of Biopharmaceutical Statistics*, **21**, 846–859.

Wang, W. and Peng, J. (2015) A step-up test procedure to find the minimum effective dose. *Journal of Biopharmaceutical Statistics*, **25**, 525–538.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D. and Hochberg, Y. (1999) *Multiple Comparisons and Multiple Tests Using the SAS System*. Cary, NC: SAS Institute Inc.

Westfall, P. H. and Young, S. S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley-Interscience.

Whitney, M. and Ryan, L. (2009) Quantifying dose-response uncertainty using Bayesian model averaging. In: *Uncertainty Modeling in Dose Response: Bench Testing Environmental Toxicity* (Ed. R. C. Cooke), 165–179. John Wiley & Sons, Inc.

Williams, D. A. (1971) A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, **27**, 103–117.

Williams, D. A. (1972) The comparison of several dose levels with a zero dose control. *Biometrics*, **28**, 519–531.

Williams, G. M. (1974) The direct toxicity of alpha-naphthylisothiocyanate in cell culture. *Chemico-Biological Interactions*, **8**, 363–369.

Yanagawa, T. and Kikuchi, Y. (2001) Statistical issues on the determination of the no-observed-adverse-effect levels in toxicology. *Environmetrics*, **12**, 319–325.

Zhang, J., Berntenis, N., Roth, A. and Ebeling, M. (2014) Data mining reveals a network of early-response genes as a consensus signature of drug-induced *in vitro* and *in vivo* toxicity. *The Pharmacogenomics Journal*, **14**, 208–216.

Appendix A

Validation of Fractional Polynomial Method in the Context of the Disconnect Analysis

In order to evaluate a performance of the modelling approach proposed in Chapter 8, two simulation studies were conducted. As a measure of performance, we estimated sensitivity and specificity. The specificity represents the rate of genes with no disconnect that are correctly not identified as disconnected genes (i.e. related to Type I error). The sensitivity represents the rate of truly disconnected genes being identified as disconnected (i.e. power of the method). The closer to one both quantities are, the better is the performance of the method. The first simulation study was focused on evaluation of sensitivity and specificity on the single gene expression experiment. The second study generated the data set resembling the structure of the data in the TGP and focused on the multiplicity adjustment, i.e. testing on thousands of genes simultaneously.

Table A.1: Simulation settings. The first two columns determine the type of profile and identification of the setting. Following two columns states explicitly the model used for particular setting and the values of parameters in model (8.1) (for case of $p_1 \neq p_2$), *in vitro* in upper panel of the table and *in vivo* in bottom panel. The specification of parameters p_1, p_2 is omitted if $\beta_1 = 0$ or $\beta_2 = 0$, respectively.

Polynomial	Setting	Model *in vitro*	Parameters *in vitro*
Null model	A	$Y_{ij} = 0 + \varepsilon_{ij}$	$\beta_0 = \beta_1 = \beta_2 = 0$
Linear	B	$Y_{ij} = (1 - \frac{Q}{3}) + \frac{Q}{3}D + \varepsilon_{ij}$	$\beta_0 = 1 - \frac{Q}{3},\ \beta_1 = \frac{Q}{3},\ \beta_2 = 0,\ p_1 = 1$
	C	$Y_{ij} = 0 + \varepsilon_{ij}$	$\beta_0 = \beta_1 = \beta_2 = 0$
	D	$Y_{ij} = 0 - \frac{Q}{3} + \frac{Q}{3}D + \varepsilon_{ij}$	$\beta_0 = -\frac{Q}{3},\ \beta_1 = \frac{Q}{3},\ \beta_2 = 0,\ p_1 = 1$
2nd order	B2	$Y_{ij} = 1 + \frac{Q}{50}D^2 + \frac{Q}{5}D^{-3} + \varepsilon_{ij}$	$\beta_0 = 1,\ \beta_1 = \frac{Q}{50},\ \beta_2 = \frac{Q}{5},\ p_1 = 2,\ p_2 = -3$
	C2	$Y_{ij} = 0 + \varepsilon_{ij}$	$\beta_0 = \beta_1 = \beta_2 = 0$
	D2	$Y_{ij} = \frac{2}{3} + \frac{Q}{50}D^2 + \frac{Q}{5}D^{-3} + \varepsilon_{ij}$	$\beta_0 = \frac{2}{3},\ \beta_1 = \frac{Q}{50},\ \beta_2 = \frac{Q}{5},\ p_1 = 2,\ p_2 = -3$

Polynomial	Setting	Model *in vivo*	Parameters *in vivo*
Null model	A	$Y_{ij} = 0 + \varepsilon_{ij}$	$\beta_0 = \beta_1 = \beta_2 = 0$
Linear	B	$Y_{ij} = (1 + \frac{4Q}{3}) - \frac{Q}{3}D + \varepsilon_{ij}$	$\beta_0 = 1 + \frac{4Q}{3},\ \beta_1 = -\frac{Q}{3},\ \beta_2 = 0,\ p_1 = 1$
	C	$Y_{ij} = \frac{4Q}{3} - \frac{Q}{3}D + \varepsilon_{ij}$	$\beta_0 = \frac{4Q}{3},\ \beta_1 = -\frac{Q}{3},\ \beta_2 = 0,\ p_1 = 1$
	D	$Y_{ij} = 0 + \varepsilon_{ij}$	$\beta_0 = \beta_1 = \beta_2 = 0$
2nd order	B2	$Y_{ij} = 1 - \frac{Q}{50}D^2 - \frac{Q}{5}D^{-3} + \varepsilon_{ij}$	$\beta_0 = 1,\ \beta_1 = -\frac{Q}{50},\ \beta_2 = -\frac{Q}{5},\ p_1 = 2,\ p_2 = -3$
	C2	$Y_{ij} = -\frac{2}{3} - \frac{Q}{50}D^2 - \frac{Q}{5}D^{-3} + \varepsilon_{ij}$	$\beta_0 = -\frac{2}{3},\ \beta_1 = -\frac{Q}{50},\ \beta_2 = -\frac{Q}{5},\ p_1 = 2,\ p_2 = -3$
	D2	$Y_{ij} = 0 + \varepsilon_{ij}$	$\beta_0 = \beta_1 = \beta_2 = 0$

A.1 Simulation study I: Performance of proposed method

A.1.1 Simulation settings

In the first simulation study, data were generated according to seven possible scenarios. The first setting (A in Table A.1) corresponds to the null model of no disconnect between two data sets. The mean profile of the other settings are presented in Table A.1 and shown in Figure A.1 (for choice $Q = 1.5$). They are generated either under a linear model (B, C, D) or a second order fractional polynomial (B2, C2, D2). The settings correspond to three groups described in the Section 8.4: genes with opposite direction of effect of the dose for *in vitro* and *in vivo* data (B, B2), genes with dose effect only for *in vivo* data (C, C2) and dose effect only for *in vitro* data (D, D2). For each setting, $N = 10,000$ data sets were generated.

For setting A, the data were generated under varying noise, i.e. with $\varepsilon_{ij} \sim N(0, SD^2)$, where $SD = 0.01, 0.14, 0.25, 0.5, 1, 1.5$. Additionally, the data were generated twice, once with same amount of observations per dose as original TGP data (two for *in vitro* and three for *in vivo*) and once with four observations per dose in both data sets.

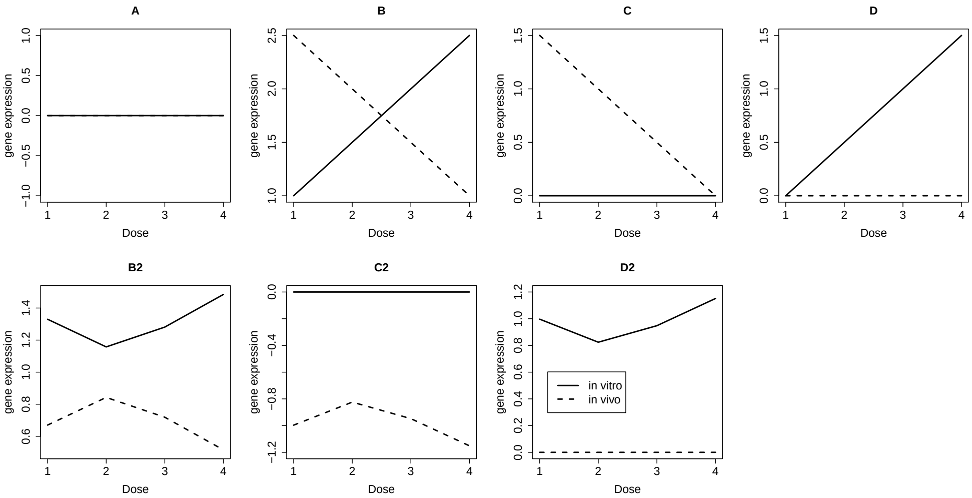


Figure A.1: The profiles used in the simulation study: means used for *in vitro* (solid line) and *in vivo* (dashed line) for the four simulation scenarios. In scenario 'A', both profiles overlap each other.

All the remaining settings (B, C, D, B2, C2, D2) were generated with value of $Q = 1.5, 2, 3$ and $\varepsilon_{ij} \sim N(0, 0.14^2)$. For settings B, C, D, the constant Q equals the fold change (as defined in Chapter 8, i.e. maximal difference of dose-specific means between the two data sets). The actual fold change for settings B2, C2 and D2 resulting from values of Q is given in Table A.4 below. The standard deviation was used as $SD = 0.14$ which approximately correspond to 75% quantile of all variances across all compounds, both for *in vitro* and *in vivo* data. The same number of observations as in the original TGP data set were used.

When the data were analysed, both test for dose-response and test for interaction were applied with level of significance 0.1. For all the settings was conducted analysis starting with *in vitro* data set, except for settings C and C2, where analysis starting from *in vivo* data set was conducted (otherwise, no disconnect would be detected, because there is no signal for *in vitro* data in C and C2).

The results for sensitivity and specificity for all scenarios are shown in Table A.2, Table A.3 and Table A.4, respectively. The specificity of separate LRTs (Table A.2) is lower than value 0.9. It is caused by the AIC procedure that selects a model with the optimal powers. The small amount of observations, especially for *in vitro* data, causes fitting more complex models than necessary. However, we can see that using both tests together (column 'Disconnect') corrects specificity of disconnect determination (given the 0.1 significance level used for testing). Additionally, a small increase of observations number per dose to $n = 4$ would improve the performance of individual tests.

The high sensitivity for LRT in case of linear model is apparent for any setting (Table A.3). The effect of fold change of one (that was considered as lowest important in our analysis) is found in all $N = 10,000$ simulated data sets. Similar pattern can be detected, when data were generated according to second order fractional polynomial models (Table A.4). The detection of disconnect is driven by dose-response detection mainly, because interaction is easily detected in all settings. For all settings, we can see high sensitivity for the values close to fold change of one which was the lowest effect of interest in our analysis and approaching maximal possible sensitivity already at fold change less than two. The higher sensitivity in setting C2 compared to D2, while having same fold change, occurs due to the dose-response effect estimated using three observations per dose *in vivo* instead of only two for *in vitro* data set. The same sensitivity for model B2 and C2 is given by fact that their dose-response profile *in vitro* is parallel, i.e. the LRT tests the same mean structure.

Table A.2: Specificity of the methodology for single experiment. The first columns determine the type of profile, number of observations per dose and the value of SD that was used to generate noise. For number of observations, TGP denotes same setting as in original data set and $n = 4$ four observations for both data sets. Following three columns show specificity of LRTs. Third column shows specificity of LRT for significance of dose-response relationship *in vitro*. Fourth column shows specificity of LRT for significance of interaction, i.e. projection of optimal fractional to both data sets. Last column represents test for disconnect, i.e. gene being significant in both LRTs for dose-response and interaction. All tests use significance level 0.1. Results of each row are based on mean of 10,000 experiments.

Profile	n	SD	*in vitro* dose-response	Projection of FP	Disconnect
A	TGP	0.01	0.8	0.81	0.9
		0.14	0.8	0.81	0.9
		0.25	0.8	0.81	0.9
		0.50	0.8	0.81	0.9
		1.00	0.8	0.81	0.9
		1.50	0.8	0.81	0.9
	$n = 4$	0.01	0.85	0.93	0.98
		0.14	0.85	0.93	0.98
		0.25	0.85	0.93	0.98
		0.50	0.85	0.93	0.98
		1.00	0.85	0.93	0.98
		1.50	0.85	0.93	0.98

Table A.3: Sensitivity of the methodology for single experiment with underlying linear model. The first two columns determine the type of profile and true underlying effect. Following three columns show sensitivity of LRTs. Third column shows sensitivity of LRT for significance of dose-response relationship *in vitro* (B, D) or *in vivo* (C). Fourth column shows sensitivity of LRT for significance of interaction, i.e. projection of optimal fractional to both data sets. Last column represents test for disconnect, i.e. gene being significant in both LRTs for dose-response and interaction. All tests use significance level 0.1. Results of each row are based on mean of 10,000 experiments.

Profile	Fold change	Dose-response	Projection of FP	Disconnect
B	0.75	0.995	1.000	0.995
	1.00	1.000	1.000	1.000
	1.50	1.000	1.000	1.000
C	0.75	1.000	1.000	1.000
	1.00	1.000	1.000	1.000
	1.50	1.000	1.000	1.000
D	0.75	0.995	1.000	0.995
	1.00	1.000	1.000	1.000
	1.50	1.000	1.000	1.000

Table A.4: Sensitivity of the methodology for single experiment with an underlying second order fractional polynomial model. The first two columns determine the type of profile and true underlying effect. Following three columns show sensitivity of LRTs. Third column shows sensitivity of LRT for significance of dose-response relationship *in vitro* (B2, D2) or *in vivo* (C2). Fourth column shows sensitivity of LRT for significance of interaction, i.e. projection of optimal fractional to both data sets. Last column represents test for disconnect, i.e. gene being significant in both LRTs for dose-response and interaction. All tests use significance level 0.1. Results of each row are based on mean of 10,000 experiments.

Profile	Q	Fold change	Dose-response	Projection of FP	Disconnect
B2	1.50	0.969	0.609	1.000	0.609
	2.00	1.293	0.796	1.000	0.796
	3.00	1.939	0.976	1.000	0.976
C2	1.50	1.151	0.802	1.000	0.802
	2.00	1.313	0.951	1.000	0.951
	3.00	1.636	0.999	1.000	0.999
D2	1.50	1.151	0.609	1.000	0.609
	2.00	1.313	0.796	1.000	0.796
	3.00	1.636	0.976	1.000	0.976

A.2 Simulation study II: Multiplicity adjustment

The second simulation study mimics the structure of the TGP experiment. In total, $M = 6,000$ genes were generated to create one data set. Half of them followed the null model for both *in vitro* and *in vivo*. The other half exhibits clear dose-response effect *in vitro* and disconnect between *in vitro* and *in vivo*. Specifically, the model used for *in vitro* was second order polynomial model

$$Y_{ij} = \frac{2}{25}D^2 + \frac{2}{5}D^{-3} + \varepsilon_{ij}.$$

The same model was used *in vivo*, disconnect was caused by increasing mean in second dose by one and decreasing mean in last dose by 0.5. The mean profile of the setting is displayed in left panel of Figure A.2. Such setting induce the fold change of one that was the minimal fold change of interest in our analysis. The $SD = 0.14$ was used, as in first simulation study, and the number of observations per dose was same as in TGP data set. Within whole data set of M genes, LRTs for dose-response and interaction were applied for each gene. The resulting p-values were adjusted for multiplicity using Benjamini-Hochberg procedure to control false discovery rate (BH-FDR). The disconnect of the gene was determined based on significance in both of the LRTs, with level of significance 0.1 used. The sensitivity and specificity was computed as amount of correctly identified genes from both categories (null model and true disconnect). The whole procedure was repeated for $N_2 = 1,000$ simulated data sets, computing sensitivity and specificity for each of them.

ROC curve of one data set is shown in middle panel of Figure A.2, showing how the sensitivity and specificity changes if significance level varies. For all $N_2 = 1,000$ simulated data sets, average sensitivity and specificity were and 0.951 and 0.932, respectively. Minimal values across all 1,000 data sets were 0.930 for sensitivity and 0.915 for specificity, suggesting consistently very good behaviour of the method when multiplicity adjustment is applied. The boxplot of all the values of sensitivity and specificity for 1,000 simulated data set is shown in right panel of Figure A.2. The specificity is well controlled, always above value of 0.9, while sensitivity is maintained very high.

In summary, both simulation studies suggest very good behaviour of the method with high sensitivity and specificity for effect of interest (fold change more than one).

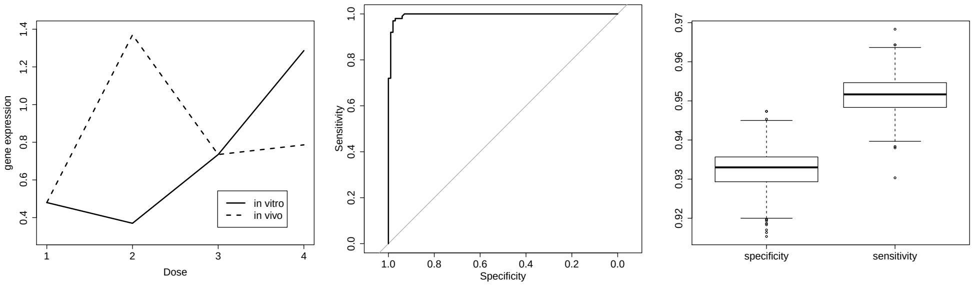


Figure A.2: Left panel: The profiles used in the second simulation study: means used for *in vitro* (solid line) and *in vivo* (dashed line). Middle panel: Sensitivity and specificity of one of the data sets when varying the significance level threshold. Right panel: Boxplot of sensitivity and specificity of all 1,000 simulated data sets.

Samenvatting

Deze thesis focust zich op dosis-respons relaties in de ruime zin. De beschreven methoden kunnen toegepast worden op ieder experiment met categorische blootstelling en een continue respons, zoals bijvoorbeeld bij de ontwikkeling van medicijnen en ecologische of economische studies. De variabelen gerelateerd aan deze blootstelling kunnen tijd, dosis, leeftijd, temperatuur enz. zijn. De natuurlijke orde is de belangrijkste eigenschap van het experiment.

De beschreven methoden in deze thesis bevinden zich op de grens van biostatistiek en statistische bio-informatica. Hoewel de focus vooral ligt op de algemene methodologisch ontwikkeling, werd het onderzoek uitgevoerd met data van hoge dimensionaliteit in het achterhoofd. De analyse uitbreiden naar data van hoge dimensionaliteit impliceert dat de analyse van een enkel experiment overgedragen dient te worden naar een situatie waarbij duizenden experimenten met dezelfde studie-opzet gelijktijdig uitgevoerd worden. In dergelijk geval is het onmogelijk om ieder experiment te evalueren door gebruik te maken van visualisatie technieken of meerdere modellen te fitten zoals typisch gedaan wordt voor een enkel experiment. Omwille hiervan zouden geautomatiseerde methoden die duidelijke beslissingsregels bieden (en bij voorkeur rekening houden met modelonzekerheid) de voorkeur moeten krijgen. Immers, in het geval van duizenden experimenten moeten multipliciteitscorrecties gebruikt worden voor een goede bescherming tegen artificiële bevindingen, veroorzaakt door toeval. Een voorbeeld van dergelijke techniek is de *false discovery rate* met multipliciteitscorrectie, een typische methode die toegepast wordt in transcriptomica.

De thesis omvat drie delen. Het eerste deel is gewijd aan de methodologische ontwikkeling terwijl de andere twee delen focussen op toepassingen binnen het domein van de bio-informatica. De structuur van de data en de modelleringsaanpak, i.e. dosis-respons experimenten en een order-restrictie modelleringsaanpak, vormen de rode draad tussen de

drie delen.

In het eerste deel van de thesis beschrijven we moderne statistische methoden op een algemene wijze zodat de methodes algemeen toepasbaar zijn. We concentreren ons zowel op de theoretische fundamenten als op de empirische evaluatie van de voorgestelde methodologie. De eigenschappen van deze methoden zijn onderzocht door uitgebreide simulatiestudies met verschillende situaties. De besproken methodologie is het Bayesiaanse variabele selectie (BVS) kader in geval van order-restrictie modellering. Het voordeel van de BVS techniek is het schatten en de model selectie gelijktijdig uitvoeren, rekening houden met onzekerheid omtrent de modellen. Deze techniek is uitgebreid met inferentie op basis van technieken die gebruik maken van het hertrekken van de steekproef. Aldus vormt het een verenigd kader zonder de noodzaak om enige post hoc methoden toe te moeten passen. Meer nog, de Bayesiaanse natuur laat toe om voorafgaande wetenschappelijke kennis in rekening te brengen wanneer ze voor handen zijn. Zoals getoond zal worden, presteren de operationele karakteristieken van de methodologie even goed als de beschikbare frequentistische technieken.

De BVS techniek wordt over verschillende hoofdstukken van het eerste deel van de thesis besproken. Hoofdstuk 2 bevat de inleiding tot het onderwerp. Hoofdstuk 3 introduceert een inferentie procedure gebaseerd op het hertrekken van de steekproef binnen het BVS kader. Model selectie en de bepaling van de minimale effectieve dosis is het onderwerp van Hoofdstuk 4. De robuustheid van de inferentie, de selectie en de schatting ten opzichte van de specificatie van de priors is onderzocht in Hoofdstuk 5. Daarenboven worden de model complexiteit en model eigenschappen gedefinieerd en geanalyseerd binnen het BVS modelleringskader in Hoofdstuk 5. Tot slot behandelt Hoofdstuk 6 in detail de opzet van de simulaties uit vorige hoofdstukken en toont bijkomende simulatie resultaten.

Het tweede deel van de thesis focust zich op de analyse van een bepaalde databank. Het doel is de ontwikkeling van de workflow om complexe data sets van meerdere bronnen te analyseren en er kennis uit te extraheren. In plaats van nieuwe methodologie te ontwikkelen, is het de bedoeling om gekende en gevalideerde methoden op een nieuwe en efficiënte wijze te gebruiken. Hoewel de aandacht gevestigd wordt op de analyse van een bepaalde databank, is het mogelijk om de workflow te veralgemenen naar gelijkaardige problemen binnen het onderzoeksdomein.

De studie die geanalyseerd wordt in het tweede deel is een grote toxicogenomische databank. Twee analyse kaders worden gepresenteerd en ieder focust van een andere visie op het translationeel onderzoek. In de eerste analyse ligt de interesse in de identificatie van genen die op dezelfde wijze reageren in twee gerelateerde datasets. Dit in tegenstelling tot de tweede analyse, waar de interesse ligt bij de identificatie van genen die sterke verschillen tonen tussen twee datasets. Beide groepen van genen zijn interessant voor verschillende onderzoeksvragen en hun identificatie zorgt voor lichtjes verschillende statistische problemen. Hierdoor variëren de gebruikte methodes van order-restrictie dosis-respons modelleringstechnieken tot de fractionele polynomen die de aanname van monotoniciteit tot op bepaalde hoogte versoepelen. De biclustering en de visualisatie van de data wordt gebruikt om interessante patronen in de data bloot te leggen. Als gevolg van de resultaten leggen we een sterke nadruk op de interpretatie van de resultaten en de identificatie van kleine interessante groepen, dit terwijl we de grote omvang van de data in rekening brengen. Het is belangrijk in het achterhoofd te houden dat beide analyses verkennende gereedschappen zijn die starten van algemene onderzoeksvragen en leiden tot een verzameling van genen. De resulterende genen blijken gewenste eigenschappen of een relatie tot de respons te bezitten, maar door de verkennende natuur van de algoritmes, dient wetenschappelijke kennis bekeken te worden en bijkomende bevestigende experimenten uitgevoerd te worden om de bevindingen te evalueren. De studie toont hoe statistische technieken succesvol toegepast kunnen worden op grote data van meerdere bronnen met uitdagende interpretatie.

De analyses van de toxicogenomische projecten worden in twee hoofdstukken gepresenteerd. In Hoofdstuk 7 wordt gezocht naar de genen die vertaalbaar zijn van *in vivo* rat naar *in vitro* mens data. In Hoofdstuk 8 worden genen met verschillende effecten over platformen, d.w.z. *in vitro* rat en *in vivo* rat, geïdentificeerd.

Tijdens het onderzoekswerk gerelateerd aan het PhD project werden grote inspanningen gedaan om data analyse technieken te voorzien voor de wetenschappelijke gemeenschap. De software ontwikkeling gebeurde in `R` (R Core Team, 2014), wegens zijn hoge kwaliteit, brede beschikbaarheid van hulpmiddelen en de vrije beschikbaarheid van `R`. In het derde deel van de thesis presenteren we twee `R` pakketten. Het eerste `R` pakket, `ORCME`, wordt gepresenteerd in Hoofdstuk 9, waarmee men order-restrictie clustering voor microarray experimenten kan uitvoeren, het kader dat typisch gebruikt wordt in de verkennende fase van de data analyse. Het pakket is beschikbaar in de Comprehensive `R` Archive Network (CRAN, Hornik, 2012) bewaarplaats en de boogde gebruikers zijn wetenschappers met minstens een basis kennis van `R`. Het tweede pakket `IsoGeneGUI`, geïntroduceerd in Hoofdstuk 10, is anderzijds geïmplementeerd als een Grafische Gebruikers Interface en is beschikbaar in Bioconductor voor een bredere gemeenschap van wetenschappers werkend op biostatistische problemen. De punt-en-klik natuur van het pakket maakt het bruikbaar voor wetenschappers met zeer beperkte kennis van `R`.