Diese Dissertation haben begutachtet:

....

Technische Universität Wien
Vienna University of Technology

DISSERTATION

Bayesian foundations for improving robustness and reliability of computational biological inference

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

o.Univ.-Prof. Dipl.-Ing. Dr.techn. Klaus Felsenstein
E107, Institut für Statistik und Wahrscheinlichkeitstheorie

und Dipl. Ing. Dr.techn. Peter Sykacek
H793 Vienna Science Chair of Bioinformatics

eingereicht an der Technischen Universität Wien
bei der TU Wien Fakultät für Mathematik und Geoinformation

von

Dipl.-Ing. Alexandra Posekany
Matrikelnummer: 0225018
Sieveringer Straße 30A, 1190 Wien

Wien, im Mai 2013

Bayesian Foundations for improving robustness and reliability of computational biological inference

Im Verlauf der letzten Jahrzehnte haben sich "high-throughput" Technologien im Bereich der biologischen und medizinischen Forschung etabliert. Diese Methoden haben zu einem Anstieg der Menge an rechnergestützten Methoden, um mit diesen Daten umzugehen, geführt. Allgemeine verwendete Annahmen der statistischen Modellbildung sind für diese Daten oft nicht erfällt, Stichprobengrößen sehr klein verglichen mit der Anzahl der zu schatzenden Parameter, während die Beobachtungen meist stark schwanken und sich alles andere als normalverteilt verhalten. Diese Herausforderung wird im Rahmen dieser Dissertation aufgegriffen.

Um mit den komplizierten Gegebenheiten der biologischen Daten umzugehen, insbesondere mit Microarraydaten, werden hierarchische Bayesmodelle konzipiert. Verschiedene Ansätze und Methoden der Robusten Bayes Statistik werden angewandt, um mit diesen Daten umzugehen. Ein hierarchischer Modellansatz hat auch den Vorteil, einen zusätzlichen Grad der Robustheit gegenüber der a-priori Verteilung und Wahl der Hyperparameter zu bieten. Um mit der Herausforderung the "overdispersion" umzugehen, werden Student t Verteilung und Mischverteilungen von t und Normalverteilungen in Betracht gezogen, um Robustheit in Bezug auf die Likelihood Funktion zu bekommen.

Die Modelle werden als Markov Chain Monte Carlo Algorithmen implementiert und mit entsprechenden Methoden aus dem Gebiet auf Konvergenzverhalten geprüft. Die biologischen Ergebnisse, die mittels dieser komplexen Ansätze gewonnen werden, werden mit existierenden Methoden aus der Bioinformatik verglichen. Darüber hinaus gehende biologische Schlussfolgerungen und Interpretationen werden ebenfall im Bereich der Bioinformatik evaluiert und auf Sinnhaftigkeit geprüft.

During the past decades high-throughput technologies have been established in biological and medical research. These methods have led to an increase in computatonal approaches to deal with their data. In addition, the data poses a challenge for data analysis. Assumptions made for general approaches in statistical modelling are often not fulfilled, sample sizes are very small compared to the number of variables to estimate, while the observations are overdispersed and the noise is behaving far from Gaussian. This challenge is met in this thesis.

For dealing with the complicated situation of biological data, in particular coming from microarrays, hierarchical Bayesian models are designed. Various ideas and Methods of Bayesian Robustness are applied for dealing with the difficult situation at hand. The hierarchical model has the advantage of providing an additional degree of robustness regarding the choice of priors and model parameters. For approaching the challenge posed by overdispersion, student's t distributions and mixtures of student's t and normal distributions are considered to gain robustness with respect to the likelihood.

The models are implemented and tested as computationally intense Markov Chain Monte Carlo sampling algorithms which are sanity checked by appropriate methods from this field. The findings gained by these more sophisticated methods are compared with existing approaches in the field of bioinformatics. Sanity checks regarding biological conclusions and interpretation of the results are gained by applying bioinformatical methods.

Acknowledgement

Not only to honour a long-lasting tradition, but also to 'reward' all those who contributed on one way or another to the completion of this work, they will be mentioned in the following. First, I would like to read the roll of honour for my comrades in arms on the battle field of bioinformatical research, German, Ulli, Pawel, Smriti, Alex, Brian, Anais, Thomas, Nancy and Nadine. Bravely we stood together to face challenges of biology, modelling and computation. Mentioning computation, I would like to thank all the nameless IT guys who suffered from my expertise in crashing laptops, desktop machines and high performance computing infrastructure all the like. Thanks to Peter for valuable advise and providing me with experience which shall be help me in my further life. Thank you KF, for maintaining the detailed balance required for the convergence of this thesis' work to this final stage. Also, my colleagues in the field of statistics, Gregor, Angela, Lars, Stefan and Laura, who provided me with valuable moral support and advice shall not go unmentioned, as little as SFS whose expertise has already guided me long before meeting in person.

Friends and foes are part of life. While foes shall be unmentioned here, some friends have become outstanding in moral support and other competences. My fool-proof proof-reader and expert for almost everything regarding English language, Bernadette, is the first to mention. But also Angela, Yumi, Eva, Sonja and Tesi have deserved their place here, distracting me and listening my endless rants about life, the universe and everything. Last, but not least my enduring family who showed all the patience in the world with my rantings and depressions about things far beyond their imagination and understanding. My humble gratitude is also deserved by WWTF and FWF without whose monetary support for scientific projects this work would not have been realised. Finally, those deserving who I forgot to mention in person, forgive my forgetfulness and accept my thanks.

Thank you, all of you!

Alex Posekany

Contents

Chapter 1

Introduction

The core of this thesis lies in developing and applying different approaches towards robust inference in a fully Bayesian setting to bioinformatical questions. Bayesian robustness considerations have their origin in frequently occurring criticism of the subjective choice of prior distributions. Starting from this, every part of the Bayesian model has become the focus of robustness considerations: prior, likelihood and loss function. Our goal was improving bioinformatical inference based on well-founded statistical theory, which we achieved by designing Bayesian models tailored for the specific type of challenges in the field of bioinformatics. On the one hand, we developed an approach towards likelihood robustness which can be linked with the idea of model selection. On the other hand, we designed a scheme for robust mixture models, which allow the identification of systematically outlying values.

Bioinformatics indisputably is a field with great need for statistical input in order to perform proper analyses of highly complex data. This research area has evolved rapidly during the past two decades alongside with high-throughput methods for biological measurements and the availablity of the necessary computing power. Such measuring methods include large projects like the human genome project which has revived and revolutionised the gene sequencing methodology as well as small platforms like microarrays, which could not be excluded from modern day biological research or medical diagnostics any more. High-throughput methods leading to gigabytes, if not terabytes of data, create new challenges for researchers in the field, especially since the number of variables can be a thousand times the number of samples. Microarrays in particular are well-known for their complicated and highly over-dispersed noise behaviour. However, a systematic analysis of their underlying structure has not been conducted before Posekany 2009 and Posekany et al. 2011 which is treated in chapter 5.1.

The importance of microarray technology for research and application has generated a plethora of sophisticated methods specifically tailored to analyse microarray data. One typical assumption in statistical data analysis is considering data to be (approximately) normally distributed. This assumption is implicit to many methods proposed for microarray data analysis including methods based on t tests (Baldi and Long 2001; Tusher et al. 2001), linear models (Smyth 2005) and Bayesian approaches (Ibrahim et al. 2002; Zhao et al. 2008; Bae and Mallick 2004; Ishwaran and Rao 2003). Even nowadays, standard approaches for the analysis of microarray data, such as limma (Smyth 2005), still assume normally distributed data even though these are sophisticated enough to use a Bayesian model structure. For almost 15 years multiple t-tests have been the gold standard to compare methods in microarray gene expression analysis with. Furthermore, the errors made by the completely unfitting normal distribution assumption have previously been unknown.

Only recently, several investigations by the bioinformatical community have cast doubt on the correctness of the Gaussian distribution assumption. By testing for Gaussianity, Hardin and Wilson 2009 found that microarray data does not follow a normal distribution at all. In fact, the observed over-dispersion manifests in a large number of outlying values, which can have considerable influence on the inference results. Both the cost of individual measurements and the possibility of outlying data points being caused by biological processes rule out that such samples get removed. The latter suggests that these values must carefully be taken into account, as excluding outlying values or including them based on incorrect distribution assumptions could falsify the resulting biological findings. Statistical techniques for determining the differential expression of genes which account for such outliers have for example been introduced by Tusher et al. 2001; Haan et al. 2009; Lee et al. 2005 and Gao and Song 2005. However, using non-parametric methods replaces the restrictive assumptions linked with the Gaussian distribution with very general ones at the cost of losing some power of tests (Whitley and

Ball 2002). An alternative to non-parametric approaches for analysing over-dispersed data is using parametric mixture densities which allow modelling Gaussian noise and deviations thereof, which are required for an appropriate consideration of outliers. Such robust noise models can for example be implemented by mixtures of Gaussian distributions or t-distributions (cf. Gottardo et al. 2006).

In the ongoing discussion about robustness of noise models, Giles and Kipling 2003 employed several statistical tests to argue that microarray data are normally distributed. On the contrary, Hardin and Wilson 2009 concluded, using similar methodologies, that heavy tailed noise is more likely to be found in microarray data. Compared to these results, the series of tests conducted by Novak et al. 2006a produced the outcome that 5-15% of genes follow a non-Gaussian distribution, while the rest is normally distributed. In the present study by Posekany et al. 2011, see chapter 5.1, 15 microarray data sets were included in a systematic study of noise behaviour. To endorse our conclusions from synthetic data about the proposed model's validity, we also used the spike-in experiment proposed in Choe et al. 2005 as a more realistic test case. In addition, we analysed 14 additionaly microarray experiments covering various experimental settings, organisms and measurement platforms. As we can see, various kinds of biological data were chosen in order to assure that our conclusions are not limited by particular choices of data sets. The data include investigations of plant soil responses, drosophila sleep deprivation, primate dietary comparisons and animal liver metabolism.

While applying our approach to the analysis of microarrays' noise behaviour, we simultaneously tested the hypothesis of differential expression. Due to the irregularity and complexity of noise, investigating the error behaviour of microarray data is of great importance, as many questions are still left unanswered. Novak et al. 2006b concluded, after testing the whole data set as well as the least extreme subsets for normality, that about 80-85 % of the data are normally distributed. However, they noted that student's t distributions provide the best fit for extreme data. Recent studies by Hardin and Wilson 2009 and Posekany et al. 2011 showed the heavy-tailed distributions' superiority over the Gaussian. Therefore, we quantitatively analysed the behaviour of over-dispersed genes, which draw the whole data set, as well as the occurrence of normally distributed genes. As the observed errors may originate from the experiments' conduction or an underlying biological process, thus being crucial for the analysis, we aim towards accounting for proper noise behaviour, as we discuss in chapter 5.

Since their first use in the 19^{th} century by Pearson 1894 for modelling sizes of crabs, mixture models have developed into a common tool in statistical inference, cf. Frühwirth-Schnatter 2006 and Mclachlan and Peel 2000. With the increase of computational power, Bayesian mixture models could be fitted where inference had been impossible before. Applications of Bayesian mixture models include model-based clustering (Banfield and Raftery 1993,Wang et al. 2011) and Bayesian mixtures (Do et al. 2005,Frühwirth-Schnatter and Pyne 2011). Ideas for robustifying Gaussian mixture models led to the development of mixtures of Student's distributions (cf. Frühwirth-Schnatter 2006) and Skew-normal or Skew-t distributions (cf. Frühwirth-Schnatter and Pyne 2011). However, often part of the data suffers from outliers, while the rest is well-fitted by the usual normal distribution. The focus of this work lies on situations in which mixtures of only Gaussians or student's t distributions would likely misjudge the situation, whereas a combination of both distributions shows a better performance, as it avoids the more complex non-Gaussian distributions, if unfitting. In such cases, our model includes normal components, whenever possible instead of student's t components with high degrees of freedom and unnecessary rescaling parameters. Including student's t distribution in the inference only makes sense, if differing between Gaussian and student's t distributions carries valuable information, as making the model more complex than a Gaussian mixture model with enough components to approximate any given likelihood implies additional computational burden. In situations where more than simple density estimation is required and the weights scattered over several normal components hold no information about outlyingness, we require a single non-normal component which can be interpreted in terms of underlying technical or biological processes.

In addition to noise estimates on microarray we introduced a measure for "non-Gaussianity" to estimate the "distance" from the normal distribution regarding the 'tail weights' of the distribution which is hard to calculate for mixtures of Gaussians. Here, the common measure for the difference between normal and student's t distribution, the kurtosis, is not defined for the most interesting distributions with degrees of freedom less or equal to

4. Hence, we applied robust functions for peakedness to be able to measure the distance between the Gaussian and the student's t distribution. Furthermore, direct inference is impossible for the degrees of freedom parameters ν of the t distributions, as we included Gaussians with theoretical degrees of freedom $\nu = \infty$ into our model. However, transforming to peakedness allows inferring these parameters without identifying their components and respective weights. To conclude, our approach fulfills the two purposes of dealing with the label switching problem and measuring "non-Gaussianity" for each gene, based on the underlying mixture components. Additionally, assessing noise behaviour and scoring the influence of over-dispersion with an according non-Gaussianity measure can be more generally applied than for microarrays, the field in which we tested the approach, see chapter 7. In bioinformatical analysis, by reducing the number of considered genes by identifying the 'possibly errorprone' or 'possibly interesting' ones due to different noise behaviour researchers could profit from this approach.

In order to give a concise overview of this thesis' structure and contents, the topics of the individual chapters will in the following be presented. Instead of ordering the chapters chronologically, we first mention the chapters dealing with theory as well as the background of the field of application. Then, we address the two chapters dealing with the two modelling approaches and their respective goals. Chapter 2 will summarise the most important theoretical background of Bayesian statistics the reader requires to understand all later chapters in this thesis except chapter 3. In addition, the not commonly known theory of Bayesian robustness will be presented in a nutshell. In chapter 3, the required bioinformatical backgrounds necessary for understanding the application of the work presented here will be touched upon. Chapter 4 discusses the theory behind Markov chain Monte Carlo sampling in detail, in order to mathematically justify and motivate its application for the models presented in chapters 5 and 7. Chapter 6 summarises the theory as well as the application of Bayesian mixture modelling for the less informed readers. In chapter 5, the first model suggested for robustification of microarray analysis which performs model comparison based on likelihood robustness consideration, will be presented. The second model will be introduced in chapter 7 and builds upon the first model, while extending it to mixtures of heavy-tailed and normal distributions with microarray quality control as a possible goal.

Chapter 2

Bayesian basics & Bayesian Robustness

2.1 The basics of Bayesian inference

In a classical sense, statistics is based on observations and their frequency. This information is then used for exploratory, often graphical, approaches as well as inferential approaches. In the framework of model-based inference, such as regression, the Likelihood function provides a formal way of including this information into the process of data analysis. Observations, characterisied by their properties (variables) and frequencies, come from a predefined space of possible observations, the population, about which we pretendedly do not know anything except for its elements. However, this scenario is rarely the case, when conducting experiments, even of an explorartory nature. Any good experimenter has an expectation of the outcome, tested for example in classical hypothesis tests.
The principal idea behind Bayesian Statistics is involving an analyst and his or her prior ideas such that any analysis becomes 'subjective' which should represent the considered experiment far better than simply stating its population. As likelihood functions are formalised by probability distributions, so are the prior believes which are introduced via prior distributions. Combining both sources of information, as in human learning, provides the data analyst with an updated, a posteriori, view of the original ideas.

Bayes' theorem formalises this information inclusion and hereby formulates the backbone of Bayesian statistics.

Theorem 1 (**Bayes' Theorem**).

For two events A and $B \in \mathfrak{S}$, where $\mathfrak{S}$ is a σ-field, the following formal relation exists:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B|A]\mathbb{P}[A] + \mathbb{P}[B|A^C]\mathbb{P}[A^C]} \tag{2.1}$$

Formalising the idea behind conditional probabilities and expectation for the continuous case,

Definition 1 (**conditional expectation**).

The conditional expectation of random variable X given the realisation of random variable $Y = y$ is defined as

$$\begin{aligned} \mathbb{E}[X|Y = y] &= \int_{D(X)} x d\mathbb{P}[x|Y = y] && (2.2) \\ &= \int x f(x|y) dx, && (2.3) \end{aligned}$$

where $D(X)$ is the domain of X.

Theorem 2 (**Bayes' Theorem II**).

In the continuous case, the posterior distribution results from

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\int_\Theta \pi(\theta) f(x|\theta) d\theta}. \tag{2.4}$$

In the denominator marginal likelihood*, $m(x) = \int_\Theta \pi(\theta) f(x|\theta) d\theta$, appears.*

This formula already includes all necessary parts required for building a parametric Bayesian model:

- the *prior distribution* $\pi(\theta)$ which expresses the uncertainty about a model parameter θ from parameter space Θ;
- and the *likelihood function* $f(x|\theta)$ which includes the information of the observations x,

- resulting in the *posterior distribution* $\pi(\theta|x)$ which relates the observations to the information about the parameter, shedding light on its behaviour in the modelled process.

2.1.1 Hierarchical Bayesian modelling

As we can see, the whole Bayesian framework is hierarchical with variables and parameters being in different points of the hierarchy. Adding another level to the model is always possible by defining prior distributions for the hyperparameters, the parameters of the prior over the parameter. Fully bayesian hierarchical models for inferring linear models have been considered for almost half a century in theory and practice, cf. Hill 1965, Tiao and Tan 1965, Robert 2001, Gottardo et al. 2003.

Directed acyclic graphs (DAG) are frequently used in the machine learning community, cf. Bishop 2006. The following convention will be used for DAGs throughout this whole thesis:

- squares around the variable symbol are used for known or predefined parameters and the data,
- circles around the symbol signify that a posterior distribution for the parameter or hyperparameter is estimated during inference.

In order to explain this notion which will be used for illustrating and supporting the bayesian hierarchical models later on, we will explain this using a short and basic example.

Example 1. *We will take a simple normal distribution model with prior for the mean and precision parameter into account. The equation*

$$y \sim N(\mu, \lambda) \quad \lambda = 1/\sigma^2 \tag{2.5}$$

presents the likelihood model for data y which is a normal distribution with mean μ and precision λ. The we define priors for μ and λ

$$\mu \sim N(m, l), \tag{2.6}$$

$$\lambda \sim Gamma(\alpha, \beta). \tag{2.7}$$

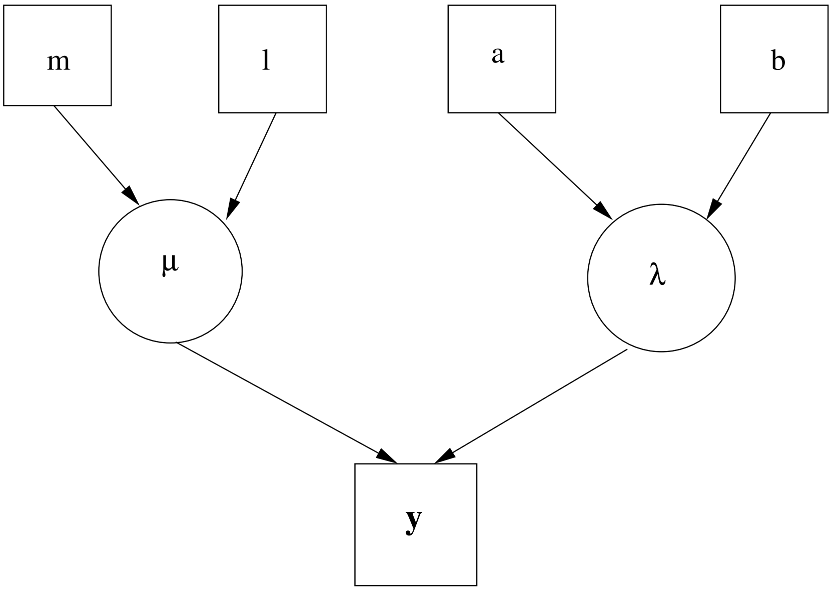


Figure 2.1: Example of a directed acyclic graph representation for the hierarchical normal distribution model with conjugate priors

This simple model can be represented graphically as in Figure 2.1

Adding another level to this hierarchy by defining e. g. a prior distribution over the parameter β in the following way

$$\beta \sim Gamma(a,b) \tag{2.8}$$

can be illustrated in the same way by adding another level in the hierarchy.

The theoretical advantage of definign hierarchical Bayesian models by adding inference over the parameters and hyperparameters lies in gaining ”objectivity” and a certain kind of ’robustness’ w. r. t. the choice of prior. Hierarchical Bayesian models are also in the center of attention of modern Bayesian statistics as they provide a flexible tool for pooling information and form the basis for computational inference methods, in particular Markov chain Monte Carlo approaches.

When defining any kind of Bayesian model, one of the most essential parts is the prior distribution’s choice. Before adding some detail on this, we take a look at two concepts playing an important role in Bayesian statistics. One is the suffiency concept, while the other is referred to as the Likelihood principle. Starting with sufficiency we define a sufficient statistic in the following way:

Definition 2 (Sufficient statistic).

For a random variable $X \sim f(x|\theta)$, we define a statistic $T(x)$ as sufficient*, if the distribution of x conditional upon $T(x)$ is independent of θ, i.e. $p(x|T(x),\theta) = p(x|T(x))$.*

The factorisation theorem provides a useful criterion for sufficiency which allows an easier check for a sufficient statistic, cf. Bernardo and Smith 2000.

Theorem 3 (Factorisation theorem).

Writing the density of x in the form

$$f(x|\theta) = g(T(x)|\theta)h(x)$$

with density g of $T(x)$ and non-negative function h is valid, iff T is sufficient.

Sufficiency forms the backbone for some fundamental principles important for any kind of statistical inference. Bayesian approaches particularly build on these principles, cf. Robert 2001 for a collection of these principles and illustrative background literature. The likelihood principle is a particular consequence of the sufficiency principle, linked by the conditionality principle.

> ***Sufficiency principle***
>
> Two observations x, y which result in the same sufficient statistic $T(x) = T(y)$ necessarily have to lead to the same inference about the parameter θ.
>
> ***Likelihood principle***
>
> The likelihood function $\ell(\theta|x)$ contains the entire information of observation x about parameter θ. If two observations x, y depend on the same parameter θ, such that there exists a constant C fulfilling the relation
>
> $$\ell_1(\theta|x) = C \cdot \ell_2(\theta|y) \ \forall\theta,$$
>
> they then contain the same information about θ and lead to the same statistical inference.

Conditionality principle

If one of two available experiments E_1, E_2 on the parameter θ is selected with probability p, the resulting inference on θ should only depend on the selected experiment.

Theorem 4 (Fundamental principles)**.** *The Likelihood principle is equivalent to the conjunction of the Sufficiency and the Conditionality principle.*

Proof. Cf. Robert 2001 □

Although Bayesian statistics has advanced far beyong its "gambling origins", which it shares with combinatorics and other parts of statistics as well, subjectivity and mainly the prior distribution introducing it still are in the line of fire by critics. The choice of a prior distribution affects the way of introducing prior information into a model, thus forming a key point of Bayesian analysis. Here, a short overview over types of prior distributions, their advantages and disadvantages is presented.

- ***Elicited prior***. A subtle and elegant way for including prior information into a model is manually creating a prior distribution, specifically based on given data. However, this comes at the disadvantage that on the one hand this method may become inconsistent rather quickly, on the other hand the laborious construction of such a prior is a waste of time if one has to feed it into a sampling algorithm later on. In addition, a reliable source of reasonable prior information is required to make this a reasonable option which rarely is the case.

- ***Natural conjugate prior***.

  Natural conjugate priors follow the straightforward way of managing a Bayesian model by choosing a prior distribution with a structure similar to the likelihood function. Thus, the prior becomes interpretable in terms of the model and allows adding previously available information, e.g. by including results from equally structured earlier experiments.

  Because of their special role in this context, we take a closer look at the exponential family of distributions. Exponential familiy distributions also play an important role in this work, since many of the distributions frequently used for statistical modelling belong to this family. Distributions belonging to this family have several advantageous properties which make them interesting for Bayesian inference (cf. Robert 2001).

Definition 3 (**Exponential Family**).

For real functions $C : \Theta \to \mathbb{R}^+, h : \mathcal{X} \to \mathbb{R}^+$ and $R : \Theta \to \mathbb{R}^k, T : \mathcal{X} \to \mathbb{R}^k$ an exponential family of dimension k *is a family of distributions with densities of the form*

$$f(x|\theta) = C(\theta)h(x)e^{R^\top(\theta)\cdot T(x)}. \tag{2.9}$$

The special case of $R(\theta) = \theta$, i.e. $R(.)$ equalling the identity $id_{\mathbb{R}^k}(.)$, is called natural exponential family.

In combination with exponential type families natural conjugate priors often describe data information using representative functions of the data, sufficient statistics. A corollary of theorem 3 is presented in the following theorem, cf. Robert 2001.

Theorem 5 (Pitman-Koopman Lemma). *If for large enough sample size there exists a sufficient statistic of constant dimension for a family of distributions $f(.|\theta)$, then this family is exponential, if the support of $f(.|\theta)$ is independent of θ.*

By placing the given restriction on the support, a line is drawn between exponential family distributions and 'quasi-exponential' distributions, such as the uniform and Pareto distribution, which share several properties typical for exponential families including the existance of constant dimensional sufficient statistics. Their support however is not independent of θ.

The converse of the Pitman-Koopman lemma that a sufficient statistic exists for any exponential family distribution is a natural property of the exponential family. Thus, one can identify a natural conjugate prior belonging to an exponential family itself, which need only be compatible with the sufficient statistic. This property justifies the additional advantage of conjugate priors that an analytical solution is always feasible, unless the model gets hierarchically structured. Updating narrows down to determining the ”new” distribution’s parameters, instead of determining the structure of the posterior. For hierarchical models these prior allows to use Gibbs updates which are favourable, as they are simple, straightforward and have a good convergence behaviour. Table 2.1 presents an overview of the most important conjugate prior settings for typical exponential family distributions which will be used for constructing the Gibbs updates later on, cf. 4.2.2. For a more detailed collection and discussion of conjugate priors consult Fink 1997.

likelihood $f(x\|\theta)$	prior $\pi(\theta)$	posterior $p(\theta\|x)$	
Normal $N(\theta, \sigma^2)$	Normal $N(\mu, \sigma_0^2)$	Normal $N(\lambda(\sigma^2\mu + \sigma_0^2 x), \lambda\sigma^2\sigma_0^2)$	$\lambda^{-1} = \sigma^2 + \sigma_0^2$
Normal $N(\mu, \theta^{-1})$	Gamma $Gamma(\alpha, \beta)$	Gamma $Gamma(\alpha + 0.5, \beta + 0.5 * (\mu - x)^2)$	
Gamma $Gamma(a, \theta)$	Gamma $Gamma(\alpha, \beta)$	Gamma $Gamma(a + \alpha, x + \beta)$	
Binomial $Bin(n, \theta)$	Beta $Beta(\alpha, \beta)$	Beta $Beta(n + \alpha, x + \beta)$	
Multinomial $Mn(n, \theta_1, \ldots, \theta_k)$	Dirichlet $D(\alpha_1, \ldots, \alpha_k)$	Dirichlet $D(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$	

Table 2.1: Overview of most important conjugate prior settings.

However, using conjugate prior has the disadvantage that information always is introduced into the model via the prior distribution. Additionally, only information fitting the structure of model and prior alike, will be passed on, any other will be disregarded. Since the conjugate prior except for the choice of its hyperparameters is predefined by the model, it is referred to as *objective*, loosing some of the subjectivity, e.g. including by an elicited prior (cf. Robert 2001). Automating the choice of prior distribution is an advantage and a nuisance. Thus, computational advantages have to be weighed carefully against disadvantages of the approach.

- ***Maximum Entropy prior***. Maximum Entropy priors base on the notion of spinning prior information into a model, based on the entropy.

Stemming from information theory, the entity of entropy measures uncertainty of data.

Definition 4 (Entropy).

For a discrete random variable X the entropy is

$$\mathcal{E}(\pi) = -\sum_{i=1}^{\infty} \log\left(\pi(x_i)\right)\pi(x_i),$$

where the sum be finite or infinite. Generalising this definition for continuous x, we define

$$\mathcal{E}(\pi) = -\int \log\left(\pi(\theta)\right)\pi(\theta)d\theta$$

The methodology of maximum entropy priors aims towards maximising the prior uncertainty, thus, being as little informative as possible in terms of entropy, while fulfilling certain side conditions. Hereby, information about certain characteristics of the prior can be included in the model, as long as they can be written as prior expectations (e.g. moments, quantiles, ...). A discrete prior maximising the entropy and prior uncertainty) can be formulated as

$$\pi^{ME}(\theta_i) = \frac{\exp\left(\sum_k \lambda_k g_k(\theta_i)\right)}{\sum_j \exp\left(\sum_k \lambda_k g_k(\theta_j)\right)}$$

Here, λ_k denote the Lagrange multipliers for optimising, while the side conditions $\mathbb{E}_\pi[g_k(\theta)] = \omega_k$ apply, where $\mathbb{E}_\pi$ refers to the first moment of the distribution π of the functions g_k of parameter θ.

In the continuous case an additional measure π_0 for reference is required.

$$\pi^{ME}(\theta) = \frac{\exp\left(\sum_k \lambda_k g_k(\theta)\right)\pi_0(\theta)}{\int \exp\left(\sum_k \lambda_k g_k(\eta)\right)\pi_0(d\eta)}$$

A maximum entropy prior, contructed in this way will by definition belong to an exponential family. Even though it allows some flexibility, while resulting in manangeable distributions, the approach has the a drawback that it is often impractical and can result in impossible parameter values, like negative "variances" ($g_1(\theta) = \theta, g_2(\theta) = \theta^2$ $\omega_1^2 > \omega_2$). Unless applying great care, moments used can become incompatible and lead to a partial rejection of available information. For example, contradictory definitions might force the analyst to drop one or more of the side conditions in order to obtain a density at all. (For details see Robert 2001)

- ***Noninformative prior***. A common problem, when performing inference, is that no prior information is available, as a prior experiment provided no compatible results or in exploratory studies. The question then becomes, how to translate 'lack of information' into a prior distribution which is a necessary part of any Bayesian model. The optimal way of formulating a single function which represents complete ignorance has not been found yet and very likely does not exist. Thus, different types of 'non-informative' priors focus on different aspects of this lack of knowledge in order to introduce "no information" partially into the model. One of the typically considered aspects is invariance to parameter transformations. Since transforming the original parameter due to easier handling of the model, e.g. standard deviation or precision instead of variance is a commonly applied means in statistics, such considerations are particularly valuable for data which results from transformation or is considered to related to a arameter of interest via transformation. Thus, a "non-informative" prior must not provide any information about the transformed parameter, when no information is available for the parameter itself. *A Catalog of Noninformative Priors* presents an extensive manual for the usage and calculation of non-informative priors.

  Jeffreys proposed a very general approach transformation invariance without introducing information. His method bases the calculation of the prior distribution on *Fisher's information matrix* with appropriate regularity conditions assumed.

$$\mathcal{I}_{ij} = \mathbb{E}\left[\frac{\partial \log\left(f(x|\theta)\right)}{\partial\theta_i}\frac{\partial \log\left(f(x|\theta)\right)}{\partial\theta_j}\right] = -\mathbb{E}\left[\frac{\partial^2 \log\left(f(x|\theta)\right)}{\partial\theta_i\partial\theta_j}\right]$$

Then the *Jeffreys noninformative prior distribution* is

$$\pi_J = [det(\mathcal{I})]^{-\frac{1}{2}}$$

which is invariant under diffeomorph parameter transformations.

Modifying Jeffrey's approach led to developing **reference priors** (Bernardo and Smith 2000). The two methods differ most notable in their way of viewing parameters: the reference prior approach distinguishes between parameters of interests and nuisance parameters, while for Jeffreys' priors all parameters are equal. Robert (Robert 2001) presented an interesting way of connecting the two approaches which also provides a constructive method for obtaining the reference prior. He modelled $x \sim f(x|\omega, \theta)$ given the multivariate parameter (θ, ω), which contains the parameter of interest θ and nuisance parameter ω. Here, the reference prior is obtained by first defining $\pi_J(\omega|\theta)$ as the Jeffreys prior of ω for fixed θ and secondly calculating the marginal distribution

$$f^*(x|\theta) = \int_\omega f(x|\omega, \theta)\pi_J(\omega|\theta)d\omega$$

The reference prior equals the Jeffreys prior $\pi_J(\theta)$, calculated with respect to the new likelihood function f^*.

Thus, the reference prior in general is an extension of the notion of Jeffreys' prior. In cases where a normal approximation of the posterior is valid, the reference prior equals the Jeffreys prior, e.g. for all continuous distributions as long as certain regularity conditions (see Bernardo and Smith 2000) are fulfilled. In the discrete case the reference prior generally equals the uniform distribution.

- ***Weakly informative priors*** Weakly informative priors were introduced to bridge the gap between informative priors and non-informative priors. They contain intentionally less information than is actually available, while defining proper prior distribution. Gelman 2006 has presented an excellent example of this kind of distribution in comparison with non-informative prior which shall be summarised here.

Example 2 (Prior distribution of the variance parameter)**.** *We look at*

Gelman's example of the 2 group hierarchical model:

$$\begin{array}{rcl} y_{ij} & \sim & N(\mu+\alpha_j,\sigma_y^2) \\ \alpha_j & \sim & N(0,\sigma_\alpha^2) \\ \sigma_\alpha^2 & \sim & \pi(\theta) \end{array} \tag{2.10}$$

- *The conditionally* conjugate *prior distribution prior distribution for σ_α^2 is of the inverse-gamma type, i. e. the distribution of its inverse, the precision, is a proper gamma distribution. This prior is most advantageous for a Gibbs updating scheme and justified, if enough data is available to outweigh the introduced prior information.*
- Non-informative *priors can be formed by improper limits of proper priors, e. g. Gamma(ϵ,ϵ) where $\epsilon \to 0$. This prior is improper and inference need not lead to a proper posterior which makes it a serious problem which should be dealt with in a sensitivity analysis.*
- *In his paper, Gelman 2006 proposes two types of weakly informative priors. One is related to the flat-tailed uniform prior, which is improper if defined on $[0,\infty)$. However, practice itself sets certain restriction that this parameter cannot become infinitely large based on finitely many finite data. Thus, a reasonable interval can be defined based on statistical expert knowledge that under certain restrictions presented by the standardisation of the data, an upper and lower bound for the uniform distribution can be found. Although, this assumes some background knowledge, this prior is not as strictly informative as the conjugate prior, adding less than what could actually be known when keeping the bounds large enough.*

Choosing a proper prior for a hierarchical Bayesian inference model has to take several aspects into account. The hierarchical structure renders some of the properties, specifically designed and built for a basic Bayesian prior-likelihood model, useless. Thus, elicited and maximum entropy priors are of limited applicability, as no reasonable prior information should be available for hyperparameters of the model parameters, which result from inference and have an interpretation. Non-informative priors lose importance as the hierarchical structure itself provides a certain degree of robustness and 'non-informativeness'. The most important criterion for prior distributions, therefore, is that they are eays to deal with, which rarely is the case for Jeffreys priors. Despite belonging to an exponential family in any case, the Maximum entropy prior need not be easy to handle. This requirement finally tips the balance in favour of the natural conjugate prior. As complex hierarchical models are usually not analytically tractable, one prefers Gibbs sampling methods as the simplest possible update and simplification of updates for Metropolis-Hastings steps, both of which which will be described in more detail in a separate chapter.

2.1.2 Bayesian Robustness

Following Berger 1994, a very brief overview of the subtypes of robust Bayesian analysis is presented. Bayesian robustness aims towards smartly choosing priors, likelihood or loss functions in such a way that the whole model becomes less sensitive to changes of other model components. The principal idea is to define a whole class of distributions instead of a single distribution in the model. For this class prior or likelihood functions are taken into account for modelling. Instead of limiting the choice to a single type of distribution, the class have a far wider range, e.g. including conjugate priors with an interval for the hyperparameters or several different distributions as possible likelihood functions. Thus, different approaches to modelling can be compared regarding their influence on the posterior. Complete 'non-informativeness' and 'lack of information' are harder to model than estimating the influence of the subjectively chosen model parts on the posterior. As noted above, it is a general problem for 'non-informative' priors to sufficiently express indifference about the parameter, which is why usually certain aspects are focussed on, e.g. transformation invariance. Walley 1991 made a good statement in that respect:

> *The problem is not that Bayesians have yet to discover the truly noninformative priors, but rather that no precise probability distribution can adequately represent ignorance.*

For example, the situation can be robustified by defining a class which includes both the natural conjugate and several non-informative and other types of prior distributions thus covering a larger range of possible model behaviour.

When working with exponential family distributions, two main problems occur during inference (see Berger 1994):

- exponential family distributions are very *sensitive against outliers*.
- *conjugate priors* have *great influence* on the inference results, if the data jars with the prior information which is implicitly introduced by its specification. This is a problem in particular if the informative choice of hyperparameters does not come from a previous study, but still is influential. This can even go as far as the prior distribution becoming more influential than the data itself, when the data is not fully compatible with the parametric model structure which is as any model only an approximation of reality. As stated by George Box Box 1979:

  > *Remember that all model are wrong; the practical question is how wrong do they have to be to not be useful.*

Bayesian statistics differs between several different concepts of robustness, discerned by their focus on the effect:

- *global robustness* compares the overall effect of change of model distributions on the parameter estimation, hypothesis testing, etc. over the total support of the model
- *local robustness* looks at the effect of change of model distributions on the parameter estimation, hypothesis testing, etc. within a suitable large neighbourhood
- *likelihood robustness* in the sense of Shyamalkumar (cf. Shyamalkumar 2000) rather compares to model selection in the sense of selecting an ”optimal”, most robust model by some definition among a *finite* set of possible models.

We will discuss the concepts of robustness consiedered in this thesis in more detail:

- **Global Robustness**

  The principal idea behind global robustness is to evaluate the overall effect of restricting the model to a class a distributions as priors or likelihood functions. The chosen class of distributions Γ is defined to contain all "reasonable" distributions. Robustness is related to the range of results, determined from all models with priors or likelihood functions in this class. This range $r(\Gamma)$ serves as an indicator whether the model is sufficiently robust. In principle, if the range $r(\Gamma)$ is not "too large" by some definition, the results are considered to be robust, cf. Berger 1994. This concept is rather vaguely defined, but very generally applicable and easy and straight forward to interpret. By choosing the thresholds for "too large" and the quantity of interest, it leaves a lot of freedom to the analyst.

  $$\begin{aligned} r(\Gamma) &= \|\overline{\psi} - \underline{\psi}\|, \\ & \quad \overline{\psi} = \sup_{\pi \in \Gamma} \psi(\pi, f), \quad \underline{\psi} = \inf_{\pi \in \Gamma} \psi(\pi, f), \end{aligned} \tag{2.11}$$

  where π represents the prior, f the likelihood function and $\psi(\pi, f)$ a decision of some kind, e. g. a point estimator from the posterior or some quantity of interest.

  For global robustness, the monotony criterion for sets and suprema and infima (2.12) holds, as written here for the one-dimensional case,

  $$\Gamma' \subseteq \Gamma \Rightarrow (\overline{\psi}' - \underline{\psi}') \leq (\overline{\psi} - \underline{\psi}). \tag{2.12}$$

  Thus, the range can always be reduced by imposing reasonable restrictions on the class Γ and hereby gaining a subset Γ' with a smaller range of results.

  The concept of global robustness applies to prior distributions, likelihood functions and loss functions alike, even simultaneously with a different set of possible functions for each of them. The downside of the approach is that by restricting the sets at will, it is always possible to fall below a certain size for the range. Keeping the balance between a suitable size of the range and the class is a delicate matter. Thus, other notions of defining ”robustness” have been considered as well.

- **Local Robustness**

  Local robustness is closely related to the notion of global robustness in the sense of looking for variation of the posterior estimates, when varying the model components, such as prior, likelihood or loss functions. However, only the local changes, in a predefined neighbourhood, are considered instead of a global measure of divergence, as is the case for global robustness. The definitions of ’suitably large’ may vary and lie in the hand of the researcher. Sensitivity analyses regarding the hyperparameters in hierarchical models usually try to validate local robustness. Here, usually an interval or neighbourhood of possible values for each hyperparameter is considered and the region of ’no considerable influence’ on the posterior distribution or estimators is determined.

- **Likelihood robustness**

  In the majority of cases Bayesian robustness consideration focus on robustifying the prior distributions. Two main reasons exist for these considerations. Firstly, since the early days of Bayesian analysis, prior distributions have been in the focus of criticism, as they form the subjective part of the method. Many statistician including Bayesian view them as the weakest link of the theory, which is why notions of ’non-informative’ priors or ideas like global robustness have been called into life. Yet, the likelihood function influences the analysis considerably by determining the way how the data will introduced into the model. However, an easy way of quantifying the actual influence does not exist, leading us to the second reason why too many considerations of likelihood robustness have been avoided: investigating the posterior robustness with respect to the likelihood is no easy task.

  Shyamalkumar 2000 was the first to propose an alternative method for approaching the challenge posed by likelihood robustness from a different direction than global and local robustness. Berger 1994 devised the original concept of global robustness to work for priors and likelihood functions alike, defining a class of distributions Γ_f from which to choose

> the likelihood function and calculating the range of results as an indication the model's robustness (see equation (2.11)). Shyamalkumar chose the way of investigating likelihood robustness by selecting the likelihood function from a *finite* class of models $M = \{M_1, \ldots, M_I\}$, which might be determined e.g. by distributions with different tail behaviour or skewness. Among these possible models one looks for the 'optimal' model to determine the most robust behaviour. Thus, his approach is far closer to model selection than the original approach by Berger, which did not ask for an optimal model. Instead, it investigated the 'influence' of the model structure and choices of parameters on results.
>
> The advantage of Shyamalkumar's method, thus, lies in its easier handling. Unlike for the determination of global infima and suprema, the complexity of calculation does not increase significantly with the increase of sample size. However, the obvious disadvantage is that only an approximation of uncertainty can be achieved, since a finite class lacks the adaptivity of a more generally defined (infinite) class.

Various approaches exist in the literature, where we briefly wish to mention only a selective few. Berger 1994 provides a detailed discussion of the then state of the art of Bayesian robustness, Berger et al. 1995 presents a collection of interesting works in the field. Wasserman 1996 very early discussed conflicts arising between improper priors, often resulting from the notion of non-informative priors, and robustness, as discussed here. Ruggeri 2010 more recently builds a bridge between non-parametric approaches, based on Dirichlet processes, and Bayesian robustness notions.

Chapter 3

Introduction to Microarray technology

The term microarray (shortened MA) does not refer to a single well-defined device of measurement. Instead, it sums up a variety of platforms which all have in common that high density assays are performed in parallel on a solid support. The basic concept is to take advantage of certain hybridisation properties of nucleic acids, when interacting with chosen complementary molecules - the *probes*, also called reporters or oligos - on a solid surface. These interactions are assumed to behave in a way that a quantitative measurement of a specific molecule of interest - the *target* or sample - can be conducted. The large scale of molecules which can be considered at once separates the microarray technology from other previous methods in biology and biochemistry, such a chromotography or thin layer electrophoresis.

In general, microarrays can be applied for any type of biochemical molecule, which fulfil the binding assumption, thus, they are also used for a broad range of applications. The most common biological substances are: tissues, proteins and DNA/RNA. Protein microarrays track interactions and activities of proteins in order to determine their function in the cell. DNA microarrays test whether parts of the DNA - *sequences* - are actively used in cells by having them react with their anti-sequence which would be located on the opposite strand of DNA in the double-helix. Tissue microarrays are tools in medical diagnosis and treatment which consider proteins, RNA or DNA molecules for a series of tests, which are performed on the patient's tissue on the small scale of the microarray.

During the past 2 decades microarrays have gained considerable influence in biological and medical investigations. From now on we refer to DNA microarrays, when simply writing microarrays. For the fabrication of microarrays a variety of technologies can be used, among them: printing with fine-pointed pins onto glass slides, photo-lithography using pre-made masks, photo-lithography using dynamic micro mirror devices as well as "ink-jet" printing, and electrochemistry on micro-electrode arrays. Some of these systems are available for creating custom microarrays in one's own laboratory. The majority of platforms used for experiments however is provided by companies specialised in microarray design and production which guarantee a minimum quality and predefined well-thought of design.

Based on the material and methods applied, two main systems of DNA microarrays are commonly available:

- **cDNA microarrays** This system is also referred to as "spotted arrays" and created by robotic spotting of genes and expressed sequence tags which have been amplified in a polymerase chain reaction where millions of copies of each gene are produced. Figure 3 visualises the typical procedure of such an microarray experiment. With the encyme reverse transcriptase the RNA is rewritten into cDNA. Each target molecule then binds to the corresponding probe on the chip providing the 'anti-sequence'. Typically, two types of DNA are compared on each microarray and then excited with laser light. The tissues are colour-coded by Cyanine3 which submits light rays in the green part of the visible light's spectrum and Cyanine5 corresponding to the red part of the spectrum. The light intensity of each part of the spectrum is then measured separately. The basic assumption of microarray analysis is that these light intensities are proportional to the original amount of a gene in the cell. Absolutely values have no meaning, as they differ too much between experiments and experimenters; only the relative differences between the light intensities can be modelled and interpreted.

- **High density oligonucleotide arrays** Single-channel (one-colour) microarrays work similar to two-colour arrays w. r. t. amplification

(PCR), and binding to its counter-part. However, now each array provides light intensity data for a single type of cells only. Again only the relative differences between the measurements of an experiment can be modelled. These platforms provide a reliable methodology and allow to easily include more than two types of tissue in the experiment design which would cause problems in the two-colour array's world which have typically been meant to compare 2 tissues only.

In order to understand the goals of bioinformatical microarray analysis, we mention some of the objectives typical for microarray studies. When possible, we demonstrate this with examples which we apply our method on, cf. Chapter 5 and 7.

- distinguishing between patients and (healthy) control persons, e. g. in testing a drug or therapy

- identification of subgroups of patients, e. g. identifying different types of carcigenous melanoma

- examination of drug response, e. g. time series of neurons within a reasonable time frame after application of an antidepressant

- comparison of alternative experimental conditions, e. g. different versions of a newly designed drug

- examination of cellular pathways, e. g. apoptosis pathway, leading to programmed cell death and other cell-cycle-relevant genes

- identification of genes for further genetic studies, e. g. identifying candidate genes for identifying cancer types or Alzheimer in medical screening

- detecting SNPs, single nucleotide polymorphisms, e. g. used for forensic analyses, evaluating mutations of cancer or other cells and determining differences between separated populations

An important aspect in conducting a microarray experiment is determining a proper experimental design which has both biological and statistical aspects. The biological experiemntal design includes the choice of platform

(e. g. oligonucleotide), probes and location of these probes, lathough the two latter aspects are typically dealt with by the company designing and producing the arrays, yet custom arrays are possible. Typically, such custom designs focus either on a specific subselection of genes or changing the probes location on the surface avoiding cross-gene effects and increase correct probe-target binding to optimise e. g. thermodynamical considerations, cf. Mückstein et al. 2012. Additional aspects of design can be die swaps to remove the known confounding effect of the die used for flourescence labelling in two-colour arrays, but also placing "control" genes on the array which are not targeted by the experiment yet meant to help detect whether something went wrong during conducting the laboratory experiment. Such an approach is the usage of "spiked in" genes which would not be present in the sample and are added with a known concentration. We will use a data set with such spiked in genes as a comparison to fully biological gene pool based data sets, cf. Table 5.4 and 7.4.1.

To understand the statistical aspects of microarray experimental design, we have to first bring to our mind the whole procedure of analysing such experimental data and its final goal, the typically gene-wise, sometimes across genes comparison of expression patterns. As the data collected from microarray experiments is usually very noisy and includes far less samples than variables, its analysis has posed an extraordinary challenge for researchers. These have to be dealt with properly before applying any kind of further data analysis or inference. In order to understand the problems occurring in microarray data analysis, we outline the process, its methods and assumptions, see also Speed 2003. The R BioConductor software, cf. Gentleman et al. 2004, provides packages which deal with all steps of analysis.

* The first step in microarray analysis is **image processing**, where the light intensity on the pictures which result from the experiment has to be evaluated properly. The most important issue here comes with placing the proper grid over the image in order to uniquely identify the spots on the scanned image. Firms provide segmentation algorithms and gridding software, appropriate for their microarrays. *Flagging*, i. e. removing or marking poor-quality and low-intensity features is mainly based on "rule of thumb" criteria.

* As a second step, the **data** is **processed**, such that global or local background light is subtracted from the measurements to make them at all meaningful, as the light intensity varies in each experiment even between arrays and is adapted to the required setting. Only after *background subtraction*, spot intensities and intensity ratios can be determined in a reasonable way. Then, a method for global or local *normalisation* is applied to the intensities or intensity ratios. This normalisation method is on the one hand extremely influential, as it determines in which way the data are made more homogeneous, before applying some kind of statistical inference.

  The classical microarray error model used for normalisation is the **additive - multiplicative error model**,

$$z = a + \epsilon + b \cdot x \cdot exp(\eta), \tag{3.1}$$

  where ϵ and η are errors of different sources, originating e. g. from biological variation or laboratory work. Based on this error model different normalisation methods are introduced, cf. Huber et al. 2004:

  - For *Quantile normalisation*, the observations of different biological states are transformed in such a way that their quantiles match.
  - For *Loess normalisation*, a smooth loess function, based on the R software system, is fit and the values of different biological systems are corrected in such a way that they would then lie on a straight regression line.
  - *Variance stabilising normalisation* (vsn) stands for a statistically well-founded transformation which is based on two assumptions for the errors:

    1. antisymmetry: $h(z_1, z_2) = -h(z_2, z_1) \qquad \forall z_1, z_2$
    2. homoskedasticity: $Var(h(z_1, z_2)) = const.$ independent of z_1, z_2

    Then Huber et al. 2002 suggested the function

$$h(z_1, z_2) = arsinh\left(\frac{z_1 - a}{\beta}\right) - arsinh\left(\frac{z_2 - a}{\beta}\right) \qquad \beta = \frac{\sigma_a b}{\sigma_b}, \qquad a, b \in \mathbb{R} \tag{3.2}$$

for stabilising the sample variance.

* The typical inference problem at hand, when looking at microarray data is determining differences in the mean gene expression between different biological settings, e. g. healthy people and cancer patients. The still standard approaches for comparison are classical tests, such as the t-test for two-colour arrays, frequentist or empirical Bayesian ANOVA or the non-parametric equivalent Mann–Whitney test. All methods have been specifically tailored to microarray data sets taking into account multiple comparisons or dimension reducing approaches, such as factor analysis or clustering. The BioConductor package limma, cf. Smyth 2005, provides an empirical Bayes software implementation in R for fitting linear models to microarray data.

  The results of such expression analyses can provide input for further approaches which consider the interaction between genes in an experiment, most prominently gene interaction networks in computational biology. Further such methods include the KEGG approach or the *gene ontology* system which builds hierarchical tree structures over biological effects. Within each level of the tree genes are assigned to specific groups defined e. g. by their function in the cell. A Fisher's exact test is then applied to each ontology in order to determine whether the number of differentially expressed genes detected in the experiment belonging to this ontology compared to the total number of differentially expressed genes differs from the marginal proportion defined by the total number of genes of this organism belonging to the ontology among all considered genes of the organism.

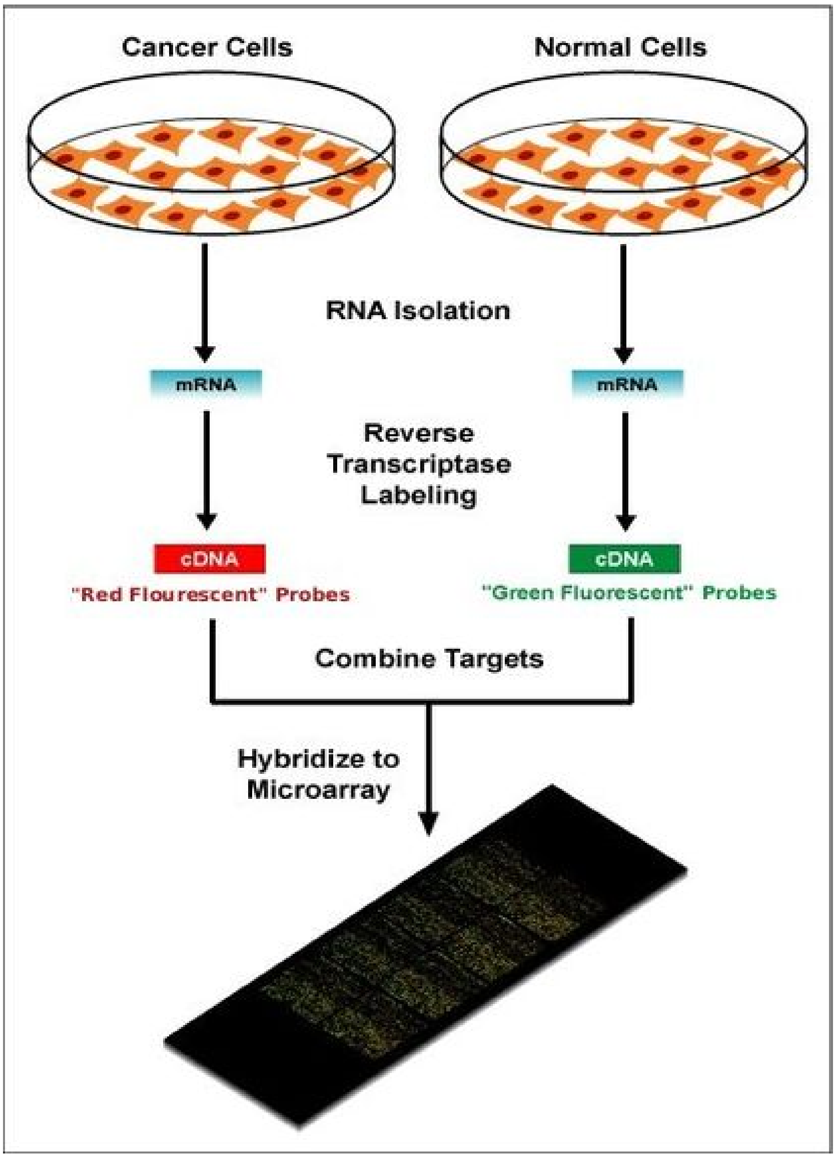


Figure 3.1: Typical procedure of two-colour microarray experiments comparing healthy against cancer tissue. (source:*DNA Microarray*)

Chapter 4

MCMC schemes

4.1 Background of the Markov Chain Monte Carlo methodology

Markov Chain Monte Carlo (MCMC) methods unite 2 principal concepts:

1. Markov Chains and

2. Monte Carlo integration.

MCMC approximates expectations with means of random draws from a given distribution combined with Markov chains which under certain regularity conditions simulate draws from a stationary distribution. The essential background knowledge required for understanding MCMC simulation is presented in the following chapter. The following definitions and theorems are required for discussing the theoretical behaviour of the presented algorithms.

4.1.1 Monte Carlo integration

Classical Monte Carlo integration has its origins in computational physics. Basically, the method was designed to calculate terms like moments of random variables or functions thereof. Given observations $(X_1, \ldots, X_n)$, generated from the known density $f(.)$, the *empirical average*

$$\overline{h_n} = \frac{1}{n}\sum_{i=1}^{n} h(x_i) \tag{4.1}$$

represents a valid approximation to

$$E_f[h(X)] \;=\; \int_{\mathcal{X}} h(x)f(x)dx. \qquad (4.2)$$

According to the Strong Law of Large Numbers, we may rest assured that $\overline{h_n}$ converges almost surely to $E_f[h(X)]$.
Under the additional condition that $h^2(.)$ has a finite expectation under f, one can actually assess the speed of convergence. In order to construct *convergence tests* for the Monte Carlo methodology, this property will be of great importance. The reason for this lies in the possibility of calculating the variance of h_n, which is given by

$$V(\overline{h_n}) = \frac{1}{n}\int_{\mathcal{X}} (h(x) - E_f[h(X)])^2 f(x)dx.$$

For practical purposes we focus our attention on its empirical estimator

$$v_n \;=\; \frac{1}{n^2}\sum_{i=1}^{n}(h(x_i) - \overline{h_n})^2.$$

According to classical theory, the Monte Carlo estimator converges against the true value in the following sense:

$$\frac{\overline{h_n} - E_f[h(X)]}{\sqrt{v_n}} \;\dot{\sim}\; N(0,1).$$

Thus, the Monte Carlo methodology provides an unbiased estimator for large enough sample sizes ($n \to \infty$). As real life restricts us to finite sample sizes, the question regarding what sample size is large enough has to be answered with smart empirical approaches. Some of those approaches of relevance for the Markov Chain Monte Carlo methodology will be discussed in 4.3.

4.1.2 Markov Chain theory

Constructing Markov Chain with the unknown distribution as its stationary distribution is the goal of MCMC. For the following theorems and definitions we will mainly rely on Robert and Casella 1999 which provides an excellent overview over Monte Carlo based sampling techniques. We will not discuss time series in general here, for an introduction on these confer Kendall and Keith 1990, but we focus instead on the Markov process theory required for our sampling algorithms. First, we require the definition of the specific kind of stochastic process which gives the MCMC methodology its name, Markov chains.

Definition 1 (**Markov Chain**).

In a discrete time setting, a Markov Chain *is a stochastic process where the sequence of random variables* $X_1, X_2, \dots$ *fulfils the* Markov property*,*

$$P[X_{n+1} = x | X_n = x_n, \dots, X_0 = x_0] \quad = \quad P[X_{n+1} = x | X_n = x_n], \quad (4.3)$$

i.e. the probability of choosing a value x *at time point* $n+1$ *given all previous values* $x_0, \dots, x_n$ *is independent of all but its precursor* $X_n = x_n$*.*

More generally in continuous time, defining Markov chains requires the concept of its transition kernel, the function which determines how the chain moves between its states.

Definition 2 (**Transition kernel**).

A transition kernel *is a function* $\mathcal{K}$ *defined on* $\mathcal{X} \times \mathcal{B}(\mathcal{X})$ *such that*

1. $\mathcal{K}(x, .)$ *is a* **probability measure** $\forall x \in \mathcal{X}$: *, i.e. for every fixed value* x *of the state space* $\mathcal{X}$ $\mathcal{K}(x, .)$ *is a function operating on the Borel sets which assigns a probability (depending on x) to every set of* $\mathcal{B}(\mathcal{X})$*;*

2. $\mathcal{K}(., A)$ *is a* **measurable function** $\forall A \in \mathcal{B}(\mathcal{X})$:*, i.e. for every fixed set* A *the function* $\mathcal{K}(., A)$ *operates on the state space* $\mathcal{X}$ *and is measurable.*

Based on the transition kernel, the Markov process has the following probability of choosing a value x_{n+1} *from set* A *of*

$$P[X_{n+1} \in A | X_n = x_n] \quad = \quad \int_A \mathcal{K}(x_n, dx) \qquad (4.4)$$

For discrete space $\mathcal{X}$ the transition kernel is a matrix with entries

$$\mathcal{K}_{x,y} = P[X_{n+1} = y | X_n = x] \qquad x, y \in \mathcal{X}$$

In order to explain the interpretation of the transition matrix we look at the probability of reaching the set A when starting from x:

$$P_x(X_1 \in A) = K(x, A).$$

Based on the notion of transition kernels, a *Markov chain* is defined as a sequence of random variables which fulfils the Markov property in (4.3) and allows us to express the probability of reaching a point in the set A when coming from $X_n = x_n$ as

$$P[X_{n+1} \in A | X_n = x_n] = \int_A \mathcal{K}(x_n, dx).$$

A Markov chain is called *homogeneous*, if the distribution of $(X_{t_1}, \ldots, X_{t_k})|X_{t_0} = x_{t_0}$ is the same as the distribution of $(X_{t_1-t_0}, \ldots, X_{t_k-t_0})|X_0 = x_0$, i.e. the distribution does not change, if it is shifted by a fixed amount of time t_0.

As the transition kernel describes the movement of the chain in one time step, the wish to take several steps at once occurs naturally. In a recursive way, the movement of the chain is described by

$$\mathcal{K}^1(x, A) := \mathcal{K}(x, a) \tag{4.5}$$

$$\mathcal{K}^n(x, A) := \int \mathcal{K}^{n-1}(y, A)\mathcal{K}(x, dy) \forall n > 1 \tag{4.6}$$

An important property of transition kernels is reflected in the Chapman-Kolmogorow equations, which provide convolution formulas of the type $\mathcal{K}^{n+m} = \mathcal{K}^n \star \mathcal{K}^m$ for the kernel for n+m transitions.

Lemma 1 (Chapman-Kolmogorow equations).

For every $(m, n) \in \mathbb{N}^2, x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$,

$$\mathcal{K}^{m+n}(x, A) = \int_{\mathcal{X}} \mathcal{K}^n(y, A)\mathcal{K}^m(x, dy).$$

Proof. see Meyn and Tweedie 1996, p. 67 □

These equations describe the probability to reach a set A in $m + n$ steps when starting from point x. Here, ones accounts for all interim values y

which can be reached from x in m steps and allow to reach A using m steps. This idea will be important for the notion of irreducibility.

A property of importance for dealing with Markov Chains is the time at which the chain enters a certain set for the first time. The term of stopping time formalises this concept.

Definition 3 (**Stopping time**).

For $A \in \mathcal{B}(\mathcal{X})$ the stopping time *is the first index n for which the chain lands in A, i.e.*

$$\tau_A \quad = \quad \inf\{n \geq 1 : X_n \in A\},$$

with $\tau_A = +\infty$, if X_n never reaches A ($X_n \notin A \ \forall n$).

In association with a set A, the number of passages *of (X_n) in A is defined as*

$$\eta_A = \sum_{n=1}^{\infty} I_A(X_n)$$

The probability of return to A in a finite number of steps*, $P[\tau_A < \infty]$, is related to both terms.*

The stopping time is an important property due to its link with other characterisitics which we will discuss in the following. It also allows to differ between the notion of weak and strong Markov property.

Definition 4 (**Weak** and **Strong Markov property**).

Let X_n be a Markov chain, h a function and $(x_0, \ldots, x_n)$ a sample from the chain and μ a probability measure, which we call the initial distribution. If the chain X_n has the weak Markov property*, then the conditional expectation of following events given the sample is independent of the sample, i. e.*

$$E_\mu[h(X_{n+1}, X_{n+2}, \ldots)|x_0, \ldots, x_n] \quad = \quad E_{x_n}[h(X_{n+1}, X_{n+2}, \ldots)], \quad (4.7)$$

assuming the expectations exist.

Given a probability measure μ and a stopping time τ which has to be finite *almost surely, if the chain has the* strong Markov property*, then*

$$E_\mu[h(X_{n+1}, X_{n+2}, \ldots)|x_0, \ldots, x_n] = E_\tau[h(X_{n+1}, X_{n+2}, \ldots)], \quad (4.8)$$

assuming the expectations exist.

There are several properties to look at, when studying a Markov chain's sensitivity to its initial conditions. Among the most important ones is irreducibility. This notion describes the possibility of reaching any point in the state space in a finite number of steps, independent of the chain's starting point.

Definition 5 (Irreducibility).

For discrete state space $\mathcal{X}$, a chain is irreducible, if all states communicate, i.e.

$$P_x[\tau_y < \infty] > 0, \quad \forall x, y \in \mathcal{X}.$$

Given an auxiliary measure μ, the Markov chain (X_n) with transition kernel $\mathcal{K}(x, y)$ is μ-irreducible *if every set $A \in \mathcal{B}(\mathcal{X})$ which is not a null set ($\mu(A) > 0$) can be reached from every point $x \in \mathcal{X}$ in a finite number of steps n, i. e. there exists an n such that*

$K^n(x, A) > 0 \;\; \forall x \in \mathcal{X} \;\; \Leftrightarrow \;\; P_x[\tau_A < \infty] > 0.$

It is strongly μ-irreducible *if n=1 for all μ-measurable sets A.*

In the following theorem, cf. Robert and Casella 1999, we summarise certain properties that are sufficient in order two imply irreducibility of a chain:

Theorem 2 (Irreducibility of (X_n)).

The Markov chain (X_n) is μ-irreducible if and only if for every $x \in \mathcal{X}$ and every $A \in \mathcal{B}(\mathcal{X})$ such that $\mu(A) > 0$, one of the following properties holds:

- *The chain can reach any set A starting from any point x in a finite number of steps n, $\exists n \in N : \mathcal{K}^n(x, A) > 0$*

- *The expected number of passages is positive, $E[\eta_A] > 0$;*

- *The resolvant is positive,* $\mathcal{K}_\varepsilon(x,A) := (1-\varepsilon)\sum_{i=0}^{\infty}\varepsilon^i\mathcal{K}^i(x,A) > 0$ *for an* ε *with* $0 < \varepsilon < 1$

Proof. cf. Meyn and Tweedie 1996, p. 87 □

Among all probability measures with respect to which a chain is irreducible, one is of special interest, namely the maximal irreducibility measure. For the ***maximal irreducibility measure*** ψ the chain is ψ-irreducible and ψ dominates all other measures μ for which (X_n) is μ-irreducible; $\mu \ll \psi$. Further theoretical statements provide even constructive methods of determining the maximal irreducibility measure ψ through a candidate measure, cf. Robert and Casella 1999.

Of high theoretical relevance, though little practical value is the definition of atoms and small sets.

Definition 6 (**Atom**).

The Markov chain (X_n) *has an* atom $\alpha \in \mathcal{B}(\mathcal{X})$ *if there exists an associated nonzero measure* ν *such that*

$$\mathcal{K}(x;A) = \nu(A) \quad \forall x \in \alpha, \forall A \in \mathcal{B}(\mathcal{X})$$

If the chain is ψ*-irreducible, the atom is called* accessible *if it is not a null set w. r. t. the maximal irreducibility measure (*$\psi(\alpha) > 0$*).*

By definition atoms require kernels which are constant on a set A of positive measure. Such a notion is too strong a requirement for general Markov chains. Thus, the term of small sets is introduced which does not restrict the kernel to reach every set A in a single step with a given 'minimum' probability. Here, we require only that such a 'minimum' probability of reaching a set A exists for a positive number of steps.

Definition 7 (**Small Set**).

A set C *is* small *if there exist* $m \in \mathbb{N}\backslash\{0\}$ *and a nonzero measure* ν_m *such that*

$$\mathcal{K}^m(x,A) \geq \nu_m(A) > 0, \quad \forall x \in C, \forall A \in \mathcal{B}(\mathcal{X})$$

Thus small sets are sets which guarantee that any set $A \in \mathcal{B}(\mathcal{X})$ can be reached in a given number of steps m with positive probability, bounded from below by some measure $\nu_m(A)$ of A.
To demonstrate this idea, we consider an irreducible Markov chain on a finite set X. For a such a chain there always exists a finite number n of steps such that the chain can reach any Borel set A with positive probability. We set $\mathcal{K}^N(x, A) =: \nu_N(A)$ for the maximum number of steps $N \in \mathbb{N}$, which exists, as we only have a finite number of states in X. Due to the Chapman-Kolmogorow equations in Equation (1) this provides us with a valid definition for such a bounding measure. Thus, the set X itself is a small set. This example also shows the connection between the notion of small sets and the irreducibility property of a Markov chain.

Another relevant property of Markov chains is *periodicity.*

Definition 8 (Periodicity).

A state x has period d, if the number of steps required to return to state x is always a multiple of d, i.e.

$$d = gcd\{n \geq 1 : P[X_n = x|X_0 = x] > 0\}$$

(gcd denotes the greatest common divisor).
If the chain is irreducible, implying that all its states communicate, there can only be one value for the period.
An irreducible chain is aperiodic*, if it has period d=1.*

Irreducibility describes a chain's ability to move through the parameter space by ensuring that the chain will enter every set. Yet, this property is too weak to guarantee that a set will also be visited ‘often enough’. Wishing the chain to return to a state infinitely often in infinite time leads us to the property of recurrence, which can be viewed in a discrete setting as a ’guarantee of a sure return’.

Definition 9 (Recurrence of a state).

In a finite state-space $\mathcal{X}$, a state $x \in \mathcal{X}$ is transient*, if the average number of visits to x when starting from x, $E_x[\eta_x]$, is finite. If this is not he case, i.e. $E_x[\eta_x] = \infty$, the state x is* recurrent.

These properties of single states apply to the whole chain, if the chain if irreducible, which follows from the Chapman-Kolmogorow equations. This means for any pair $(x, y) \in \mathcal{X}^2$ that the expected number of visits to y when starting from x $E_x[\eta_x] = \infty$. Furthermore, for any Markov chain the following definition applies

Definition 10 (**Recurrence of a Markov chain**).

A Markov chain (X_n) *is* recurrent *if*

1. *there exists a measure* ψ *such that* (X_n) *is* ψ*-irreducible and*
2. $\forall A \in \mathcal{B}(\mathcal{X})$ *with positive measure,* $\psi(A) > 0: \quad E_x[\eta_A] = \infty \;\; \forall x \in A$

the chain is transient *if*

1. (X_n) *is* ψ*-irreducible*
2. $\mathcal{X}$ *is transient, i.e. all states in* $\mathcal{X}$ *are transient.*

In general, one can come up with the following classification result that recurrence and transience are dichotomous properties for ψ-irreducible chains, cf. Meyn and Tweedie 1996.

Theorem 3 (**Recurrence of a** ψ**-irreducible chain**).

A ψ*-irreducible chain is either recurrent or transient.*

This is a direct result of the Chapman-Kolmogorow equations that recurrence or transience is not a property of a single state but the chain itself. Once a single state is recurrent, all other states visited by the chain due to irreducibility inherit the same property. A more rigid property for guaranteeing to reach any state ‘often enough’ is Harris recurrence. On the one hand, it requires an infinite average number of visits for every small set, thus, implying the same limiting behaviour of the chain for almost every starting value. On the other hand, it applies to all states, providing a global property of the chain.

Definition 11 (**Harris recurrence**).

A set A is Harris recurrent*, if the chain almost surely returns to A an infinite number of times,* $P_x[\eta_A = \infty] = 1 \;\; \forall x \in A$.

The chain X_n is Harris recurrent, *if there exists a measure ψ such that (X_n) is ψ-irreducible and for every set A with positive measure, $\psi(A) > 0$, A is Harris recurrent.*

Two main results can be deduced from the notion of Harris recurrence.

Theorem 4 (Harris recurrence of (X_n)).

If every Borel set $A \in \mathcal{B}(\mathcal{X})$ has a finite stopping time almost surely, $P_x[\tau_A < \infty] = 1 \ \forall x \in A$, then the number of returns is infinite almost surely, $P_x[\eta_A = \infty] = 1 \ \forall x \in \mathcal{X}$, and the Markov chain (X_n) is Harris recurrent.

Proof. cf. Robert and Casella 1999, p.222 □

Theorem 5 (Harris recurrence of ψ-irreducible chains).

If (X_n) is a ψ-irreducible Markov chain with a small set C such that $P_x[\tau_C < \infty] = 1 \ \forall x \in \mathcal{X}$, then (X_n) is Harris recurrent.

Proof. see Meyn and Tweedie 1996, p. 206 □

The idea behind this theorem is that if a ψ-irreducible chain can independently of its starting point reach a small set in a finite number of steps given that such a set exists, it can by definition of the small set reach any other set A in a finite number of steps with positive probability. Meyn and Tweedie 1996 provide a discussion and additional theorems and proofs about recurrence and Harris recurrence.
An even higher level of stability of a chain X_n is reached if its marginal distribution becomes independent of the chain index n, which implies that for X_n and X_{n+1} a common probability distribution π exists such that $X_n \sim \pi, X_{n+1} \sim \pi$. This notion leads us to the following definitions and results.

Definition 12 (Invariant measure, positivity, stationary distribution).

A σ-finite measure π is invariant *for the transition kernel $\mathcal{K}(.,.)$ as well as for the respective chain if*

$$\pi(B) \quad = \quad \int_{\mathcal{X}} \mathcal{K}(x, B)\pi(dx), \quad \forall B \in \mathcal{B}(\mathcal{X})$$

When there exists an invariant probability measure for a ψ-irreducible chain, the chain is positive *recurrent. Recurrent chains without such a finite invariant measure are called* null recurrent.
The invariant measure π is referred to as stationary distribution *if π is a probability measure, as in that case $X_0 \sim \pi$ implies $X_n \sim \pi\ \forall n$. Such a chain is* stationary in distribution.

The following theorem clarifies the connection between positivity and recurrence.

Theorem 6 (**Positive recurrence**)**.**

If the Markov chain X_n is positive, it is also recurrent.

Proof. cf. Robert and Casella 1999, p.224 □

Kac's theorem, for details and proof see Meyn and Tweedie 1996 p. 235, is a rather classical result on irreducible Markov chains in a discrete state-space. Basically, it states that in case of existence of the stationary distribution, this stationary distribution is defined by

$$\pi_x = (E_x[\tau_x])^{-1}$$

An implication of this result is that $(E_x[\tau_x])^{-1}$ is the eigenvector associated with the eigenvalue 1 of the corresponding transition matrix. This result can also be generalised for the continuous case. A further implication of this result is the following theorem which is also important for justifying the MCMC method.

Theorem 7 (**Uniqueness of the invariant measure**)**.**

If (X_n) is a recurrent chain, there exists a invariant σ-finite measure which is unique up to a multiplicative factor.

Proof. cf. Meyn and Tweedie 1996, p. 236. Follows directly from Kac's theorem. □

Without the guarantee of uniqueness the whole setting of MCMC sampling would be rendered useless, as it depends on draws from this stationarity distribution. Without this result one could never be sure that the stationary

distribution is the 'correct' one given that the chain reaches stationarity at all.
The stability property of stationarity of a chain is related to another property, its reversibility. This notion generally states that the dynamics of the chain is not influenced by the direction of time. More formally this means

Definition 13 (**Reversibility**).

A stationary Markov chain (X_n) is reversible if the distribution of X_n conditionally on $X_{n+1} = x$ is the same as the distribution of X_n conditionally on $X_{n-1} = x$.

Tightly linked to reversibility is the detailed balance condition.

Definition 14 (**Detailed Balance Condition**).

A Markov chain with transition kernel $\mathcal{K}(.,.)$ satisfies the detailed balance condition if there exists a function f satisfying

$$\mathcal{K}(y,x)f(y) = \mathcal{K}(x,y)f(x) \quad \forall(x,y)$$

This definition provides us with a sufficient, although not necessary condition for f to be a stationary measure associated with a transition kernel $\mathcal{K}$ (and its respective Markov chain). A more general statement links this condition with the notion of reversibility.

Theorem 8 (**Detailed Balance Condition, reversibility**).

If a Markov chain with transition kernel $\mathcal{K}$ satisfies the detailed balance condition with π a probability density function, the following statements hold true:

1. *The density π is the invariant density of the chain.*

2. *The chain is reversible.*

Proof. To proof (1), we consider a measurable set B. For this set the detailed balance condition implies

$$\begin{aligned}\int_{\mathcal{X}} K(y,B)\pi(y)dy &= \int_{\mathcal{X}}\int_{B} K(y,x)\pi(y)dxdy = \\ &= \int_{\mathcal{X}}\int_{B} K(x,y)\pi(x)dxdy = \int_{B}\underbrace{\int_{\mathcal{X}} K(x,y)dy}_{=1}\pi(x)dx\end{aligned}$$

As the existence of the invariant density π is shown, reversibility follows directly from inserting π into the detailed balance condition. □

A natural candidate for the limiting distribution is the (unique) invariant distribution. In order to make a statement about this limiting distribution, a sufficient condition for (X_n) is required under which X_n is asymptotically distributed according to π. Among the many possible conditions, which one can place on the convergence of the distribution P_n of X_n, the most fundamental and important is that of ergodicity.

Definition 15 (**Ergodicity**).

For a Harris positive chain (X_n), with invariant distribution π, an atom α is ergodic *if*

$$\lim_{n\to\infty} |\mathcal{K}^n(\alpha,\alpha) - \pi(\alpha)| = 0$$

The total variation norm provides a useful statement about convergence. Here, this norm is defined as, cf. Meyn and Tweedie 1996 and Robert and Casella 1999:

Definition 16 (**Total Variation norm**).

The total variation norm *of a measure μ is used:*

$$\|\mu\|_{TV} = sup_{g:|g|\leq 1}\left|\int g(x)\mu(dx)\right| = sup_{A\in\mathcal{B}(\mathcal{X})}\mu(A) - inf_{A\in\mathcal{B}(\mathcal{X})}\mu(A) \quad (4.9)$$

which is a special case of the more general norm $\|.\|_h$

$$\|\mu\|_h = sup_{g:|g|\leq h}\left|\int g(x)\mu(dx)\right| \quad (4.10)$$

The metric which is induced by the total variation norm is defined as

$$\|\mu_1 - \mu_2\|_{TV} = \sup_A |\mu_1(A) - \mu_2(A)|$$

Definition 17 (Geometric and uniform Ergodicity).

A chain is geometrically h-ergodic*, if for a non-negative real-valued function M with $E_\pi[|M|] < \infty$ and $0 < r_h < 1$*

$$\|K^n(x,.) - \pi\|_h \leq M(x) \cdot r_h^n$$

Robert and Casella 1999 state that $M(x) = \sum_{n=1}^{\infty} r^n \|K^n(x,.) - \pi\|_h$.

A chain is uniformly ergodic*, if for constants $M > 0$ and $0 < r < 1$*

$$\sup_x \|K^n(x,.) - \pi\|_{TV} \leq M \cdot r^n$$

Among several statements, made about convergence under these conditions, the most important ones are:

Theorem 9 (Positive recurrence and convergence in the Total Variation norm).

If Markov chain (X_n) is positive recurrent and aperiodic with transition kernel $\mathcal{K}(.,.)$ and there existis an ergodic atom α, then

$$\lim_{n\to\infty} \|\mathcal{K}^n(x,.) - \pi\|_{TV} \quad = \quad 0 \;\; \forall x \in \mathcal{X}$$

Proof. see Meyn and Tweedie 1996, p. 315 □

Theorem 10 (Harris recurrence and convergence in the Total Variation norm).

If the Markov chain (X_n) is Harris positive and aperiodic, then

$$\lim_{n\to\infty} \left\| \int \mathcal{K}^n(x,.)\mu(dx) - \pi \right\|_{TV} \quad = \quad 0 \;\; \forall x \in \mathcal{X}$$

for every initial distribution μ.

Proof. see Meyn and Tweedie 1996 p. 322 □

The main result on which the theory of MCMC simulation is based is the ergodic theorem.

Theorem 11 (Ergodic theorem).

If (X_n) has a σ-finite invariant measure π, the following two statements are equivalent:

1. *If $f, g \in L^1(\pi)$ with $\int g(x)d\pi(x) \neq 0$, then*

$$\lim_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n} f(X_i)}{\frac{1}{n}\sum_{i=1}^{n} g(X_i)} = \frac{\int f(x)d\pi(x)}{\int g(x)d\pi(x)}$$

2. *The Markov chain (X_n) is Harris recurrent.*

Proof. see Robert and Casella 1999, p. 242 □

This section is intended to specify how certain properties of the Markov chain lead to conclusions about its behaviour. Figure 4.1 provides a graphical overview of the main properties required for Markov Chain Monte Carlo convergence considerations. The implications of these properties in the setting of different sampling methods, especially regarding convergence, are essential for the whole theory behind the Markov Chain Monte Carlo methodology. However, practical implications are very limited, as all these notions ranging from irreducibility to convergence assume an infinite number of draws which will never be reached in practice. This section therefore forms a theoretical backbone which serves as justification of the presented algorithms.

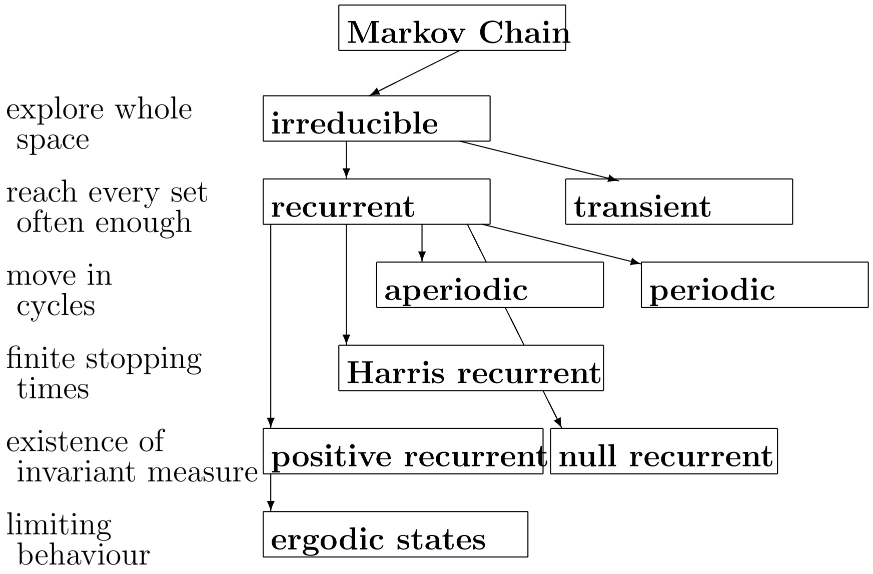


Figure 4.1: Visualisaton of basic Markov chain properties. An arrow marks the direction of becoming more specific. A property at the tip of the arrow implies the property at its shaft.

4.2 Overview of some important sampling methods

The principal idea behind any MCMC method is to obtain samples from a posterior distribution without calculating this distribution explicitly, since hierarchical Bayesian models often lead to analytically intractible posteriors. MCMC sampling schemes aim for constructing an ergodic Markov chain with stationary distribution ξ in order to acquire samples from that distribution. As described in the field of Monte Carlo integration, moments and sample-based estimators can be calculated.

We differ between several basic kinds of samplers, required for this thesis, additional ones do exist, but are not dealt with in this work:

- The **Metropolis-Hastings** sampler is the most universal sampling scheme.

- The **Gibbs** sampler presents the most commonly used, simple to understand and straightforward to calculate and implement method.

- The **Reversible Jump** sampler and the **Birth and Death** sampler provide approaches for dealing with varying parameter sizes.
- Hybrid samplers combine at least 2 of the sampling approaches listed above making them more generally applicable for hierarchical models.

4.2.1 Metropolis Hastings Sampler

The aim of the Metropolis Hastings sampler is drawing from the objective ***target density*** ξ. These draws are realised via an auxiliary conditional distribution $q(.|.)$ of a proposed value given the 'old' value. This ***proposal density*** should be either easy to simulate from or symmetric (i.e.$q(x|y) = q(y|x)$) so that it cancels out in the acceptance probability. Then, the Metropolis-Hastings sampler works according to the following scheme, described in Table 4.1. If the ratio of target and proposal function in Equation (4.11) is increased for the proposal compared to the old value, then the value is accepted for sure. Otherwise, if the ratio decreases, the

- For $t = 0$: take starting value x_0
- $t > 0$:
  1. generate proposal $Y_t \sim q(y|x^{(t-1)})$
  2. Either
     move to the proposed value Y_t with probability $\alpha(x^{(t-1)}, Y_t)$ or
     stay at the old value $x^{(t-1)}$ with probability $1 - \alpha(x^{(t-1)}, Y_t)$

     where $\alpha(x, y) = \min \left\{ \frac{\xi(y)}{\xi(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}$ is the *acceptance probability.*

     The transition kernel of the Metropolis-Hastings sampler is

$$\mathcal{K}(x, y) = \alpha(x, y) q(y|x) + (1 - \int \alpha(x, y) q(y|x) dy) \delta_x(y) \qquad (4.11)$$

Table 4.1: Generic Metropolis-Hastings sampling algorithm

proposal is accepted with probability α.

This approach is illustrated by the following simple example.

Example 3. *Metropolis-Hastings Sampler for student's t distribution*

Consider a non-central student's t-distribution model with known degrees of freedom ν and scale 1.

$$\begin{aligned} X &\sim t_\nu(\theta, 1) \\ f(x, \theta) &\propto (\nu + (x - \theta)^2)^{-\frac{\nu+1}{2}} \end{aligned}$$

To keep this example simple, we choose a flat prior for θ: $\pi(\theta) \propto 1$, and the proposal distribution is standard normal $N(0,1)$. Given 1 sample of x (adding more samples would result in a product of the above likelihood function), $\theta^{(t-1)}$ and the proposal ζ drawn from $N(0,1)$ the acceptance probability for run $t \geq 1$ would be:

$$\alpha(\theta^{(t-1)}, \zeta) = \left(\frac{\nu + (x - \zeta)^2}{\nu + (x - \theta^{(t-1)})^2}\right)^{-\frac{\nu+1}{2}} \frac{\exp\left(-\frac{1}{2}(\theta^{(t-1)})^2\right)}{\exp\left(-\frac{1}{2}\zeta^2\right)}$$

for any proposed value of θ that stays within the parameter's support. Proposals outside the support of the target density are necessarily rejected.

In order to draw conclusions about properties of the chains necessary for convergence, certain conditions are required for the functions, defining the Metropolis-Hastings acceptance probability and transition kernel. Even though the generic Metropolis-Hastings algorithm is well-defined for any target and proposal distribution, certain *regularity conditions* are of importance for ξ to be the limiting distribution of the chain:

- The support of ξ, $supp_\xi$, shall be connected, which is not necessary for the algorithm to work, but very helpful for applications and important for irreducibility and existence of a single stationary distribution

- $\cup_{x \in supp_\xi} supp_{q(.|x)} \supset supp_\xi$, i.e. the set of all values where the target distribution ξ is not zero (i.e. its support) has to be contained in the union of the supports of all possible proposals within the support of

ξ. This condition is the minimal necessary condition for ξ to be the limiting distribution of the chain.

Theorem 12 (Detailed balance condition).

Let $(X^{(t)})$ be the chain produced by the Metropolis-Hastings algorithm (see table 4.1). For every conditional distribution q whose support includes the support of ξ the following two statements hold:

1. *the kernel of the chain satisfies the detailed balance condition with ξ.*
2. *ξ is a stationary distribution of the chain.*

Proof. The proof is straightforward and can be viewed as an example for the application of the detailed balance condition. We apply the detailed balance condition (4.9) on the kernel in equation (4.11).

$$\alpha(x,y)q(y|x)\xi(x) \quad = \quad \alpha(y,x)q(x|y)\xi(y)$$

$\alpha(x,y) = \min\left\{\frac{\xi(y)}{\xi(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}$, thus 2 cases are possible.

1. $\alpha(x,y) = 1$

   $$q(y|x)\xi(x) = \frac{\xi(x)}{\xi(y)}\frac{q(y|x)}{q(x|y)} \cdot q(x|y)\xi(y)$$

2. $\alpha(y,x) = 1$

   $$q(x|y)\xi(y) = \frac{\xi(y)}{\xi(x)}\frac{q(x|y)}{q(y|x)} \cdot q(y|x)\xi(x)$$

$$(1 - \int \alpha(x,y)q(y|x)dy)\delta_x(y)f(x) = (1 - \int \alpha(y,x)q(x|y)dx)\delta_y(x)f(y)$$

Both expressions equal zero if $x \neq y$, otherwise the terms on both sides are necessarily equal. □

Aperiodicity of the chain requires that with positive probability the state $X^{(t+1)}$ may be equal to $X^{(t)}$ which is equal to

$$P[\xi(X^{(t)})q(Y_t|X^{(t)}) \leq \xi(Y_t)q(X^{(t)}|Y_t)] < 1$$

The theoretical considerations above have shown us that irreducibility is a minimum requirement for recurrence and positivity and thus for any notion of 'converging' to the invariant measure, which is our ultimate goal. Therefore our first step will be to show irreducibility with respect to ξ. *Irreducibility* of the chain can already be shown using the sufficient condition of *positivity* of the conditional density q, i.e.

$$q(y|x) > 0 \quad \forall (x,y) \in supp_\xi \times supp_\xi$$

Proposing any value in the support of ξ with positive probability independent of the current point immediately implies that in a finite number of steps any set in this support can be reached, which is equal to the definition of irreducibility according to Theorem 2.
Irreducibility and existence of the invariant distribution per definitionem imply *positivity* of the chain and thus recurrence using Theorem 6.
In general it can be proven that any ξ-irreducible Metropolis-Hastings chain $(X^{(t)})$ is *Harris recurrent.* Thus it fulfils the ergodic theorem (Theorem 11). To present this result in a more formal way the following convergence theorem is formulated.

Theorem 13 (Convergence theorem for MH algorithm).

If $(X^{(t)})$ is an ξ-irreducible Metropolis-Hastings Markov chain, the following statements hold:

- *If $h \in L^1(\xi)$, then*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) \quad = \quad \int h(x)\xi(x)dx$$

- *If in addition $(X^{(t)})$ is aperiodic, then it converges in the total variation norm, i.e.*

$$\lim_{n\to\infty} \left\| \int \mathcal{K}^n(x,.)\mu(dx) - \xi \right\|_{TV} = 0,$$

for every initial distribution μ and MH-transition kernel for n steps, $\mathcal{K}^n(x,.)$.

Proof. Robert and Casella 1999, p. 274 f. □

4.2.2 Gibbs Samplers

The Gibbs sampler can be seen as a special case of Metroplis-Hastings sampler, where every draw is accepted automatically. The simple, yet excellent and straightforward to implement idea is to use the true conditional distributions associated with the target distribution in order to generate samples from that distribution.

We require to be able to simulate from the conditional distribution $\xi_i(x_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p) \quad i = 1, 2, \ldots, p$. Then $\forall t \geq 1$ given the value $x^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_p^{(t)})$ generate

$$\begin{array}{rcl} X_1^{(t+1)} & \sim & \xi_1(x_1|x_2^{(t)}, \ldots, x_p^{(t)}) \\ X_2^{(t+1)} & \sim & \xi_2(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}) \\ \vdots & & \vdots \\ X_p^{(t+1)} & \sim & \xi_p(x_p|x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)}) \end{array}$$

The transition kernel of this algorithm is

$$\mathcal{K}(x^{(t+1)}|x^{(t)}) = \prod_{j=1}^{p} \xi(x_j^{(t+1)}|x_1^{(t+1)}, \ldots, x_{j-1}^{(t+1)}, x_{j+1}^{(t)}, \ldots, x_p^{(t)})$$

Table 4.2: p-stage Gibbs algorithm

The following example illustrates the differences between Metropolis-Hastings and Gibs sampler.

Example 4. *Gibbs Sampler for bivariate normal distribution*
Let $x = (x_1, x_2)$ follow a bivariate normal distribution of the following type

$$\left.\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right| \rho \;\sim\; N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

Then the Gibbs algorithm will update in step $t \geq 1$ as follows:

$$\begin{aligned} X_1^{(t)}|x_2^{(t-1)} &\sim N(\mu_1 + \rho(x_2^{(t-1)} - \mu_2), 1 - \rho^2) \\ X_2^{(t)}|x_1^{(t)} &\sim N(\mu_2 + \rho(x_1^{(t)} - \mu_1), 1 - \rho^2) \end{aligned}$$

A single Gibbs transition can be interpreted as a special case of a single component Metropolis-Hastings move where the acceptance probability always equals 1. Thus the 2-stage Gibbs sampler inherits all properties of the Metropolis-Hastings Sampler. However this is not the case for the multistage Gibbs sampler, which can be seen as the most well-behaved example of a *hybrid sampler*. This will be described in more detail in an extra section.

4.2.3 Introduction to Reversible Jump MCMC

The method of reversible jump Markov Chain Monte Carlo (RJMCMC) was introduced by Peter Green (see Green 1995, Richardson and Green 1997), while Waagepetersen and Sorensen 2001 presented an excellent discussion of the methodology. In principle, RJMCMC provides a generalisation of the Metropolis-Hastings method in order to allow for jumps between spaces Θ_k of different dimensionality. The main trick, but also the main challenge in building the algorithm is cleverly defining a bijection (which is even a diffeomorphism) between well-constructed spaces, containing the original spaces as linear subspaces and of course have the same dimension.
Being in the current state $x = (k, \theta^{(k)})$, where k is the indicator of the model and corresponding parameter space and $\theta^{(k)} \in \Theta_k$ the respective model parameter, a move of type m is proposed which would lead to state dy with probability $q_m(x, dy)$. The acceptance probability for such a proposal move shall be α_m. The algorithm requires a reversible kernel, which means that for some invariant density π it fulfils

$$\int_A \int_B \mathcal{K}(x, dy)\pi(x)dx = \int_B \int_A \mathcal{K}(y, dx)\pi(y)dy \qquad \forall A, B \subset \Theta$$

The appropriate kernel can be written as

$$\begin{aligned}
\mathcal{K}(x,B) &= \sum_m \int_B \alpha_m(x,y')q_m(x,dy') + s(x)I_B(x) \\
s(x) &= \underbrace{\sum_m \int_{\Theta_m} q_m(x,dy')(1-\alpha_m(x,y')) +}_{\text{probability to reject proposed move m}} \qquad \underbrace{1-\sum_m q_m(x,\Theta_m)}_{\text{probability of not attempting any move}} \\
&= 1-\sum_m \alpha_m(x,y')q_m(x,\Theta)
\end{aligned}$$

The term $s(x)$ describes the probability of rejecting the proposed move m or not attempting any move at all.
The detailed balance condition requires that

$$\begin{aligned}
&\sum_m \int_A \pi(dx) \int_B q_m(x,dy')\alpha_m(x,y') + \int_{A\cap B} \pi(dx)s(x) \\
=& \sum_m \int_A \pi(dy') \int_B q_m(y',dx)\alpha_m(y',x) + \int_{B\cap A} \pi(dy')s(y')
\end{aligned}$$

Since the last term is the same for both lines it is sufficient that for each m the respective summands of the first term of both lines are equal. In order to fulfil this a symmetric dominating measure ξ_m on Θ is required and we assume that $\pi(dx)q_m(x,dy')$ has a finite density $f_m(x,y')$ with respect to this measure. Then reversibility can be shown to be fulfiled:

$$\begin{aligned}
\int_A \pi(dx) \int_B q_m(x,dy')\alpha_m(x,y') &= \int_A \int_B \alpha_m(x,y')f_m(x,y')\xi_m(dx,dy') \\
&= \int_A \int_B \alpha_m(y',x)f_m(y',x)\xi_m(dy',dx) \\
&= \int_A \int_B \alpha_m(y',x)q_m(y',dx)\pi(dy')
\end{aligned}$$

In order for the middle equality to hold the acceptance probability has to look like

$$\alpha_m(x,y') = \min\left\{1, \frac{f_m(y',x)}{f_m(x,y')}\right\} = \min\left\{1, \frac{\pi(dy')q_m(y',dx)}{\pi(dx)q_m(x,dy')}\right\} \qquad (4.12)$$

How to obtain this dominating measure ξ_m under the symmetry constraint is the most complex part of the method when moving from model k_1 to k_2. It is supposed one has proper densities $p(\theta^{(k_1)}|k_1)$ on R^{n_1} and $p(\theta^{(k_2)}|k_2)$ on R^{n_2}. The idea of Green is to embed both spaces Θ_{k_1} and Θ_{k_2} as linear subspaces in space $\mathfrak{C}_1$ and $\mathfrak{C}_2$ which have the same dimension so that the definition of a bijection is possible. Then take a look a the spaces $U_1 = \mathfrak{C}_1 \setminus \Theta_{k_1}$ and $U_2 = \mathfrak{C}_2 \setminus \Theta_{k_2}$ with dimensions $dim(U_1) = m_1$ and $dim(U_2) = m_2$ and thus $n_1 + m_1 = n_2 + m_2$. The completion of the spaces Θ_{k_i} requires simulation of the values u_i, $u_i \sim g_i(u_i)$. Let ω be the bijection $\omega : \mathfrak{C}_1 \to \mathfrak{C}_2 : (\theta^{(k_1)}, u_1) \mapsto (\theta^{(k_2)}, u_2)$. The density f will look like

$$\begin{aligned} f(x, y') &= \pi(k_1, \theta^{(k_1)})\pi_{k_1,k_2} g_1(u_1) \\ f(y', x) &= \pi(k_2, \theta^{(k_2)})\pi_{k_2,k_1} g_2(u_2) \left| \frac{\partial \omega(\theta^{(k_1)}, u_1)}{\partial(\theta^{(k_1)}, u_1)} \right| \end{aligned}$$

Thus the acceptance probability will become

$$\min\left\{1, \frac{\pi(k_2, \theta^{(k_2)})\pi_{k_2,k_1} g_2(u_2)}{\pi(k_1, \theta^{(k_1)})\pi_{k_1,k_2} g_1(u_1)} \left| \frac{\partial \omega(\theta^{(k_1)}, u_1)}{\partial(\theta^{(k_1)}, u_1)} \right| \right\}$$

To summarise the following table presents the algorithm in a more straightforward manner.

- For $t = 0$: take starting value x_0
- $t > 0$: $x^{(t-1)} = (k_1, \theta_{k_1}^{(t-1)})$
  - Select model k_2 with probability π_{k_1,k_2}
  - Generate $u_i \sim g_i(u_i) \quad i = 1, 2$
  - $(\theta^{(k_2)}, u_2) = \omega(\theta^{(k_1)}, u_1)$
  - Accept $\theta^{(k_2)}$ with probability

$$\min\left\{1, \frac{\pi(k_2, \theta^{(k_2)})\pi_{k_2,k_1} g_2(u_2)}{\pi(k_1, \theta^{(k_1)})\pi_{k_1,k_2} g_1(u_1)} \left| \frac{\partial \omega(\theta^{(k_1)}, u_1)}{\partial(\theta^{(k_1)}, u_1)} \right| \right\}$$

Table 4.3: Reversible Jump algorithm

4.2.4 Hybrid sampler

A more general setting than the multistage Gibbs sampler is often required if the conditional distribution of a variable is not explicitly available. In this case a sampling method combining Gibbs and Metropolis-Hastings updates will be required.

Definition 18 (**Hybrid Sampling algorithm**).

A hybrid MCMC algorithm is a Markov chain Monte Carlo method which utilizes several Gibbs or Metropolis-Hastings steps. Two ways of building a hybrid kernel from the kernels $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_n$ are possible:

- *a* mixture *of steps is associated with the kernel*

$$\widetilde{\mathcal{K}} = \alpha_1 \mathcal{K}_1 + \alpha_2 \mathcal{K}_2 + \ldots + \alpha_n \mathcal{K}_n$$

  (where $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ is a probability distribution)

- *a* cycle *has a kernel*

$$\mathcal{K}^* = \mathcal{K}_1 \circ \mathcal{K}_2 \circ \ldots \circ \mathcal{K}_n$$

The motivation for constructing such samplers containing not only Gibbs steps, as the multi-stage Gibbs sampler, is that Metropolis-Hastings steps can be applied in more general settings than a Gibbs step. This is especially of importance, when the conditional distributions cannot be sampled from directly there is no alternative but to deviate from the Gibbs setting.
The Hybrid sampler is built upon full conditional distributions like the Gibbs sampler. Besag and Green Besag et al. 1995 have pointed out that for any p-variate $x, x' \in supp_\xi$ and indices $I \subset \{1, \ldots, p\}$, where x_I denotes all components of x with indices in I and x_{I^C} contains the components with indices not in I

$$\begin{aligned} \xi(x_I | x_{I^C}) &\propto \xi(x) && (4.13) \\ \frac{\xi(x_I^{'} | x_{I^C}^{'})}{\xi(x_I | x_{I^C})} &= \frac{\xi(x^{'})}{\xi(x)} \qquad for\; x_{I^C}^{'} = x_{I^C} && (4.14) \end{aligned}$$

In this way full conditionals can by easily introduced into Metropolis-Hastings steps, as the acceptance probability will become

$$\alpha(x, y) = \min \left\{ \frac{\xi(y_I|x_{I^C})}{\xi(x_I|x_{I^C})} \frac{q(x_I|y_I, x_{I^C})}{q(y_I|x_I, x_{I^C})}, 1 \right\}$$

This formula also allows us to easily see the connection between the Gibbs update and a single Metropolis Hastings step which is the case of I containing just one single index. In the case of the Gibbs sampler this proposal distribution is chosen to be

$$q(y_I|x_I, x_{I^C}) = \xi(y_I|x_{I^C}) \tag{4.15}$$

independent of x_I. Obviously the acceptance probability becomes 1 independently of x and y.

Some basic properties of the individual kernels are inherited by the hybrid kernel, for example a mixture kernel is irreducible and aperiodic if at least one of the $\mathcal{K}_i$ has these properties. If one of the kernels of a cycle is irreducible and aperiodic, then the composed kernel *often* is irreducible and aperiodic as well, however there exist counterexamples showing that this is not always the case. For any composition where each component has the same stationary distribution ξ, the stationary distribution of the composition will be ξ as well.

Under rather rigid assumptions a very specialised result can be obtained (see Tierney 1994)

Theorem 14 (Uniform ergodicity of hybrid sampler).

If $\mathcal{K}_1$ and $\mathcal{K}_2$ are two kernels with the same stationary distribution ξ and if $\mathcal{K}_1$ produces a uniformly ergodic Markov chain, the mixture kernel

$$\widetilde{\mathcal{K}} = \alpha\mathcal{K}_1 + (1-\alpha)\mathcal{K}_2 \quad (0 < \alpha < 1)$$

is also uniformly ergodic.

Moreover, if $\mathcal{X}$ is a small set for $\mathcal{K}_1$ with $m = 1$, the kernel cycles $\mathcal{K}_1 \circ \mathcal{K}_2$ and $\mathcal{K}_2 \circ \mathcal{K}_1$ are uniformly ergodic.

Proof. see Robert and Casella 1999, p. 390 □

Partially collapsed sampling

Additionally, we introduce the methodology of partially collapsed Gibbs sampling (cf. Dyk and Park 2008, Park and Dyk 2009) which will be employed in this work for improving the efficiency of our updates. The original idea is to combine 2 parameters or sets of parameters of a Gibbs scheme to a larger one and perform Gibbs updates with this new set of parameters. In a more formal setting this ansatz is shown in equation 4.16, when updating 4 sets of parameters A, B, C and D, where we want to update the parameters B and C jointly.

$$\begin{array}{lll} \text{update } A & \text{based on} & P[A|B,C,D] \\ \text{update } (B,C) & \text{based on} & P[(B,C)|A,D] \\ \text{update } D & \text{based on} & P[D|A,B,C] \end{array} \tag{4.16}$$

According to Bayes's theorem, the update of the new larger parameter or set of parameters is an update of one parameter or set of parameters based on the full conditional and the other based on all parameters except for the one updated jointly.

$$P[(B,C)|A,D] \quad = \quad P[B|A,C,D] \cdot P[C|A,D] \tag{4.17}$$

In order to use this simpler update for C which is drawn only from A and D,we must not yet update any of the parameters or sets of parameters which condition on C. Updating based on non-full conditionals would destroy the Gibbs scheme, which requires the full conditional distributions (cf. Robert and Casella 1999). Thus, we can only update parameters in such a partially collapsed step, if during this updating step no other parameter has conditioned on this parameter before. Therefore, unlike the original Gibbs sampler, the order of the updates is crucial for the partially collapsed sampler.

$$\begin{array}{lll} \text{update } C & \text{based on} & P[C|A,D] \\ \text{update } B & \text{based on} & P[B|A,C,D] \\ \text{update } A & \text{based on} & P[A|B,C,D] \\ \text{update } D & \text{based on} & P[D|A,B,C] \end{array} \tag{4.18}$$

For one of the Gibbs steps the update is partially collapsed as the jointly updated parameter is integrated out. Besag (Besag et al. 1995) has described the generalisation of MCMC updates of Metropolis-Hastings or Reversible jump type based on full conditionals. As full conditionals are the only required assumption for partial collapsing, the method can be applied to all updates in the hybrid sampler.

4.3 General MCMC Convergence Analysis

As we will refer to convergence diagnositics later in this work and present some results of such convergence diagnostics, we first provide an overview over the involved notions and statistics (for more details see Cowles and Carlin 1996). Since all MCMC algorithms have to be terminated after a usually pre-specified number of samples drawn, great importance lies in determining whether one can safely assume that the obtained samples are truly representative of the underlying distribution. As these algorithms sample based on Markov chains, the obtained sample of the posterior is generally correlated due to the autocorrelation of the Markov chain. This correlation is responsible for the draws being less informative than iid draws from the stationary distribution would be. Thus, poor mixing of the chain can be the result which means exploring the stationary distribution is hindered and slowed down.

A more theoretically well-founded attempt is analysing the transition kernel itself in order to determine the required number of iterations. However, practice showed that this is not a very fruitful method, as the resulting bounds were quite loose and too large to be of practical value. Thus, practitioners commonly apply various forms of diagnostics tools to the algorithm's output in order to draw conclusions about convergence a posteriori. Based on this information of pre-runs the actual algorithm is run for a resulting number of draws, which will likely ensure convergence. Yet, the critical issue for all these diagnostics is that independent of the construction of sample size statistic, one cannot compare sample distributions to the unknown stationary distribution. The only information available are other sample distributions which originate either from different iterations or from different parts of the same chain. Therefore, many theoreticians rightfully criticise that all such diagnostics are fundamentally unsound which does not keep many people from still using them due to lack of sound alternatives.

The following diagnostics are used in this work:

- ***Raftery and Lewis diagnostics***
  The method aims towards detecting convergence and provides bounds for the variance of the estimated quantiles of the analysed parameters or functions thereof. In order to calculate the diagnostics with a given precision, a minimum number of draws, N_{min}, is required under the assumption that they are independent. Then, the method's aim is to estimate a quantile q with accuracy r, which has to be attained with probability s, $P[q - r \leq \hat{\theta} \leq q + r] \geq s$.
  The output will be the total number of iterations to be run in order to fulfil the criterion above and the number of iterations to be considered as 'burn-in', i.e. the minimum number of iterations required for the chain to approach its stationary distribution. Additionally, it provides a 'thinning number', k, which can be seen as a representation of correlation within the chain. The idea is to remove the within-chain-correlation of the chain such that the draws would be approximately i. i. d. when keeping every k-th sample of the posterior distribution and discarding all the ones in-between.

- ***Geweke diagnostics***
  The notion behind the creation of this diagnostics is to use methods of spectral analysis to assess convergence of the sampler. The main assumption is that for a MCMC process and a given function g a spectral density $S_g(\omega)$ exists for this time series that has no discontinuities at frequency 0. The spectral density describes the distribution of variance of a time series with frequency; it can be obtained as Fourier transform of the autocorrelation function. If the conditions above are fulfiled, the spectral density provides us with the asymptotic variance $S_g(0)/n$ for the estimated mean of $g(\theta)$, $\overline{g(\theta)}_n$. This is a requirement for performing a two-sample t-test given that the conditions are fulfiled under which this diagnostic approaches a standard normal distribution according to the central limit theorem. Geweke's diagnostic after N iterations is the respective test statistic when comparing the N_1 first iterations and N_2

last iterations.

$$G_N = \frac{\overline{g(\theta)}_{N_1} - \overline{g(\theta)}_{N_2}}{S^*}$$

where S^* is the asymptotic standard error of the difference.

- ***Heidelberger-Welch diagnostics***
  This diagnostic is predicated on another approach based on the usage of methods of spectral analysis for detecting nonstationarities in outputs of MCMC algorithms. This procedure allows to estimate a confidence interval of specified width for the mean if the chain does not sample from the stationary distribution already from the beginning. The test for diagnosing convergence is based on the Brownian bridge theory from which its null hypothesis is derived. The statistic is the sum of mean-centered iterates divided by the standard error. The distribution of the Cramer-von Mises statistic is then used to test the hypothesis.

- ***Autocorrelation, Partial Autocorrelation***

Definition 19 (**Autocorrelation function**).

Autocorrelation *describes the correlation between different time points of a time series. For a discrete process of length N, the autocorrelation function is defined as*

$$\widehat{R}(k) = \frac{1}{(N-k)\widehat{\sigma_\varepsilon}^2} \sum_{t=1}^{N-k} (X_t - \overline{X}_n)(X_{t+k} - \overline{X}_n)$$

Since for a Markov chain each state depends on the previous one, we expect the time points to be autocorrelated. However, autocorrelation is one of the greatest problems in MCMC sampling as the goal is obtaining (apporixmately) i. i. d. draws of the posterior in order to estimate moments based on Monte Carlo methods. Thus, the auto-correlation has to be etsimated and taken care of. *Autoregressive process* describe one way of modelling the behaviour of time series, particularly ones generated from MCMC samplers.

Definition 20 (**Autoregressive process of order p**).

$$\theta_t = \alpha_1 \theta_{t-1} + \ldots + \alpha_p \theta_{t-p} + \varepsilon_t \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

This model is a linear regression model, where the value at time t is predicted by the p previous values $\theta_{t-1}, \ldots, \theta_{t-p}$. The *partial autocorrelation* helps to estimate the order p of such a model as it estimates the correlation between X_t and X_{t-k} that has not been explained by $X_{t-1}, \ldots, X_{t-k+1}$ for all k, the maximum value k for which this partial autocorrelation is still significantly different from 0 is the autoregressive model order.

- ***Gelman-Rubin Diagnostic***
  Unlike the other methods presented here, the Gelman-Rubin diagnostic is applied to multiple chains. Basically, it can be viewed as an analysis of variance among two or more chains which ideally should have started from different even overdispersed initial values. Its goal is to find multimodality and thus determine whether at least one of the chains gets stuck at a local peak.
  Based on the empirical variances of every single chain on the one hand and all chains combined on the other hand the Gelman-Rubin diagnostic calculates a so-called ***shrinking factor***. Values of this statistic which are close to 1 point towards convergence, whereas values significantly greater than one indicate problematic behaviour.

Several of these diagnostics have been implemented in the R package coda by Plummer et al. 2006b which we will use for our analyses. We will combine some of these diagnostic tools, in order to be able to gain insight into several aspects of the chains' behaviour. The tests provided by the methods of Heidelberger-Welch and Geweke give us general insight into the occurrence of convergence. Both allow us to compare subsets of the first 50 % of draws to the second half, if the null hypothesis is not accepted immediately due to slow burn-in. If the null hypothesis of this halfwidth test is rejected, gradually the first 10%, 20%, etc. are discarded and the rest of draws is tested against the second half. If these tests fail every time, clearly no convergence has occurred

Method	quant./graph.	Theoretical basis
Raftery-Lewis	quantitative	2-state Markov chain theory
Geweke	quantitative	Spectral analysis
Heidelberger-Welch	quantitative	Brownian bridge spectral analysis
Gelman-Rubin	quant./qual.	analysis of variance within and between chains
Autocorrelation	quant./graph.	Correlation of the sample
Partial autocorrelation	graphical	from a single Markov chain

Table 4.4: Overview of diagnostic methods and some of their properties; all these methods have in common that they are generally applicable to any MCMC algorithm, only take at least one chain into account and work for univariate parameters (cf. Cowles and Carlin 1996 for more details)

and the only interesting diagnostic would be Raftery-Lewis' prediction for the estimated number of draws necessary for convergence. However the Raftery-Lewis diagnostics can be seen as both a prediction of run lengths for future draws as well as a 'sanity' check in case of convergence, if enough draws have occurred at all in order to gain sufficiently accurate estimates of the posterior distribution. If more than one chain is available comparing them using the Gelman-Rubin diagnostic is advisable as comparison of more than one chain is the only way to detect local convergence problems caused e.g. by multimodality. Plotting the first values of the autocorrelation function helps to detect problems of slow mixing and may empirically provide the number of values to be left out in order to get iid draws, if it is not too large. A check of the Markov property is possible using the partial autocorrelation function.

Chapter 5

Likelihood robustness in MA analysis

The following chapter introduces a recently developed model for investigating Bayesian robustness issues in microarray data analysis. Guided by the notion of likelihood robustness, we performed a systematic study of a variety of data sets, stemming both from biology as well as laboratory work.

5.1 Model structure

The above mentioned Bayesian hierarchical model developed by Posekany 2009 was specifically designed to investigate the robustness of error models in the bioinformatical analysis of microarrays. For this purpose, an ANOVA type linear model was linked to the investigation of the biologically relevant question of differential expression, implemented as a latent indicator variable I_g. The ansatz for this approach is the following linear model equation,

$$y_{n,g} = x_{n,g}^T \beta_g + \varepsilon_{n,g}, \qquad n = 1, \ldots, N, g = 1, \ldots, G \tag{5.1}$$

where for any given sample n and gene g the model variables represent the following biological concepts:

$y_{n,g}$ is the observed light intensity, corresponding to gene expression;

$x_{n,g}$ is a vector of the underlying design matrix indicating the biological system which a sample n belongs to;
$x_{n,g} = [\mathbb{I}(S_{n,g} = 1), \ldots, \mathbb{I}(S_{n,g} = S)]^T \in \mathbb{R}^{S \times 1}$.

$S_{n,g}$ is a factor variable, encoding the biological system that observation $y_{n,g}$ belongs to and is defined by the experiment's design. It is included in the model by the design matrix X.

X $(x_{1,g}, \ldots, x_{N,g}) =: X_g = X \in \mathbb{R}^{S \times N}$
The design matrix $(x_{1,g}, \ldots, x_{N,g}) =: X_g = X \in \mathbb{R}^{S \times N}$ is based on the dummy coding of biological systems $S_{n,g}$ w.r.t. n and is independent of gene g, as all arrays are equal thus containing the same genes.

β_g is the vector of mean expressions of a gene for the S different systems.

I_g is the biological indicator which differs between differential expression and no differential expression of a gene g.

$\varepsilon_{n,g}$ are the noise residuals.

For biological interpretation, the parameter of interest is the differential expression indicator I_g. The posterior distribution of this parameter evaluates the probability of each gene to be differentially expressed and allows ranking genes according to their biological relevance. In typical microarray experiments' analyses, we are only interested in differentially expressed genes. In the statistical model, I_g differs between between a univariate and a multivariate linear model by determining the dimension of the coefficient vector β_g. A one-dimensional parameter refers to the null hypothesis of the ANOVA model that the gene g is not differentially expressed. Here, non-differential expression are defined as all biological systems having the same mean expression.

$$\begin{aligned} I_g = 0: \quad \beta_{g,0} | I_g = 0 \quad &\sim \quad N_1(\mu_{g,0}, (\tau_{g,0})^{-1}) \\ &\qquad \beta_g = [\beta_{g,0}, \ldots, \beta_{g,0}]^T \ \in \mathbb{R}^{S \times 1} \end{aligned} \tag{5.2}$$

The alternative hypothesis we wish to consider in our model scenario proposes that gene g is not differentially expressed. If the estimated mean expression of at least one group differs from the rest, gene g is by definition differentially expressed. Here, we model the coefficient vector β_g as a multivariate vector, which contains the different estimated mean expressions for the respective groups.

$$\begin{aligned} I_g = 1: \quad \beta_g | I_g = 1 \quad &\sim \quad N_S(\mu_g, T_g^{-1}) \\ &\quad\quad \mu_g = [\mu_{g,1}, \ldots, \mu_{g,S}]^T \ \in \mathbb{R}^{S \times 1} \\ &\quad\quad T_g = \begin{pmatrix} \tau_{g,1} & & 0 \\ & \ddots & \\ 0 & & \tau_{g,S} \end{pmatrix} \in \mathbb{R}^{S \times S}. \end{aligned} \tag{5.3}$$

Posekany 2009 decided for the introduction of a hierarchical model structure due to the advantages of this approach, cf. Section 2.1.1, as well as its natural structure which is ideal for such complicated situations. Our main reasons for this choice are on the one hand robustness with respect to the choice of hyper-parameters and on the other hand the ability to specifically model those parameters on a higher level of the model which still have interpretations. An example for such a hyper-parameter with a reasonable model-inherent interpretation would be the overall differential expression probability. This probability p of any gene to be differentially expressed regarding the overall differential expression behaviour of the experiment is the hyper-parameter of the Bernoulli prior distribution for I_g.

$$I_g | p \quad \sim \quad Bin(1, p) \tag{5.4}$$

In the hierarchical structure, the probability p is updated using a Beta distribution, which is the natural conjugate prior in this setting,

$$p \quad \sim \quad Be(a, b). \tag{5.5}$$

The part of this model which is most important for the following work is the alternative noise model. The standard approach for linear ANOVA models is to assume normally distributed residuals. As robustness with respect to the assumed distribution of the observations is our main aim in this approach, we want to allow for different noise distributions. The noise distribution corresponds to the likelihood function in the Bayesian model which is the most difficult part of the model to focus on when aiming for robustness, cf. Section 2.1.2. In order to handle outlying and over-dispersed data points in a more suitable way, it is assumed that the set of considered distributions contains student's t distributions in addition to the normal distribution. We tackled the resulting challenge of selecting the most suitable error distribution not by a posteriori model comparison, but by including the inference of the models into our approach. In the following section, the details of this advantageous model comparison approach will be presented.

5.2 The Student's t error model

The general framework of the Bayesian hierarchical ANOVA model described in section 7.1 is our starting point for robustifying the noise inference. As discussed in Section 5.1, our ansatz includes Student's t-distributions as possible likelihood functions. However, student's t distribution present a moderate challenge for Bayesian modelling, as they not only have no conjugate prior distribution, but also form a likelihood function which is very difficult to handle. In order to tackle these problems, we employed a hierarchical Bayesian representation. Following Bernardo and Smith 2000, the non-central t-distribution's likelihood function can be replaced by a hierarchical structure consisting of a Normal- and a Gamma-distribution in the following way:

$$X \sim t_\nu(\mu, \sigma^2) \Leftarrow \begin{array}{l} X|\varphi \sim N(\mu, \frac{1}{\varphi}\sigma^2) \\ \varphi \sim Ga(\frac{\nu}{2}, \frac{\nu}{2}) \end{array} \tag{5.6}$$

According to (5.6) we can rewrite our model as

$$\begin{aligned} y_{n,g}|\beta_g, \nu &\sim t_\nu(x_{n,g}^T\beta_g, \tau_\varepsilon^{-1}) \Leftarrow \\ y_{n,g}|\beta_g, \varphi &\sim N(x_{n,g}^T\beta_g, (\varphi_{n,g}\tau_\varepsilon)^{-1}) \\ \varphi_{n,g}|\nu &\sim Ga(\tfrac{\nu}{2}, \tfrac{\nu}{2}) \\ \tau_\varepsilon|g, h &\sim Ga(g, h). \end{aligned} \tag{5.7}$$

The introduced auxiliary parameter $\varphi_{n,g}$ can hereby be interpreted as a rescaling factor of the normal distribution's variance such that outlying values become more probable. In the now following lemma, we prove that the marginal distribution of $y_{n,g}$ is indeed a student's t distribution, as this is a constructive proof, thus helpful in understanding the whole model approach.

Lemma 1. *The marginal distribution $m(y_{n,g}|\nu)$ of $y_{n,g}$ follows a student's t distribution with degrees of freedom ν.*

Proof.

$$\begin{aligned} p(y_{n,g}, \varphi_{n,g}|\ldots) &= \underbrace{\frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}}_{=: c_1}(\varphi_{n,g})^{\frac{\nu}{2}-1}\exp{(-\frac{\nu}{2}\varphi_{n,g})}\underbrace{\frac{\tau^{0.5}}{\sqrt{2\pi}}}_{=: c_2}\varphi_{n,g}^{0.5}\exp{(-\frac{1}{2}\tau\varphi_{n,g}(y_{n,g}-x_{n,g}^T\beta_g)^2)} \\ &= c_1c_2\underbrace{\varphi_{n,g}^{\frac{\nu+1}{2}-1}\exp{(-\varphi_{n,g}\frac{1}{2}(\nu+\tau(y_{n,g}-x_{n,g}^T\beta_g)^2))}}_{=: I(\varphi_{n,g})} \end{aligned}$$

The expression $I(\varphi_{n,g})$ shares the structure of a Gamma-distribution $Ga(a, b)$ with parameters for shape $a = \frac{\nu+1}{2}$ and rate $b = \frac{1}{2}(\nu+\tau(y_{n,g}-x_{n,g}^T\beta_g)^2)$. All that is missing here is the normalisation constant. Therefore, the marginal distribution is

$$m(y_{n,g}) = \int_0^\infty I(\varphi_{n,g})d\varphi_{n,g} = c_1c_2\frac{\Gamma(a)}{b^a}$$

After cancelling a few terms we gain

$$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\frac{\tau^{0.5}}{\sqrt{\nu\pi}}(1+\frac{\tau}{\nu}(y_{n,g}-x_{n,g}^T\beta_g)^2)^{-\frac{\nu+1}{2}}.$$

□

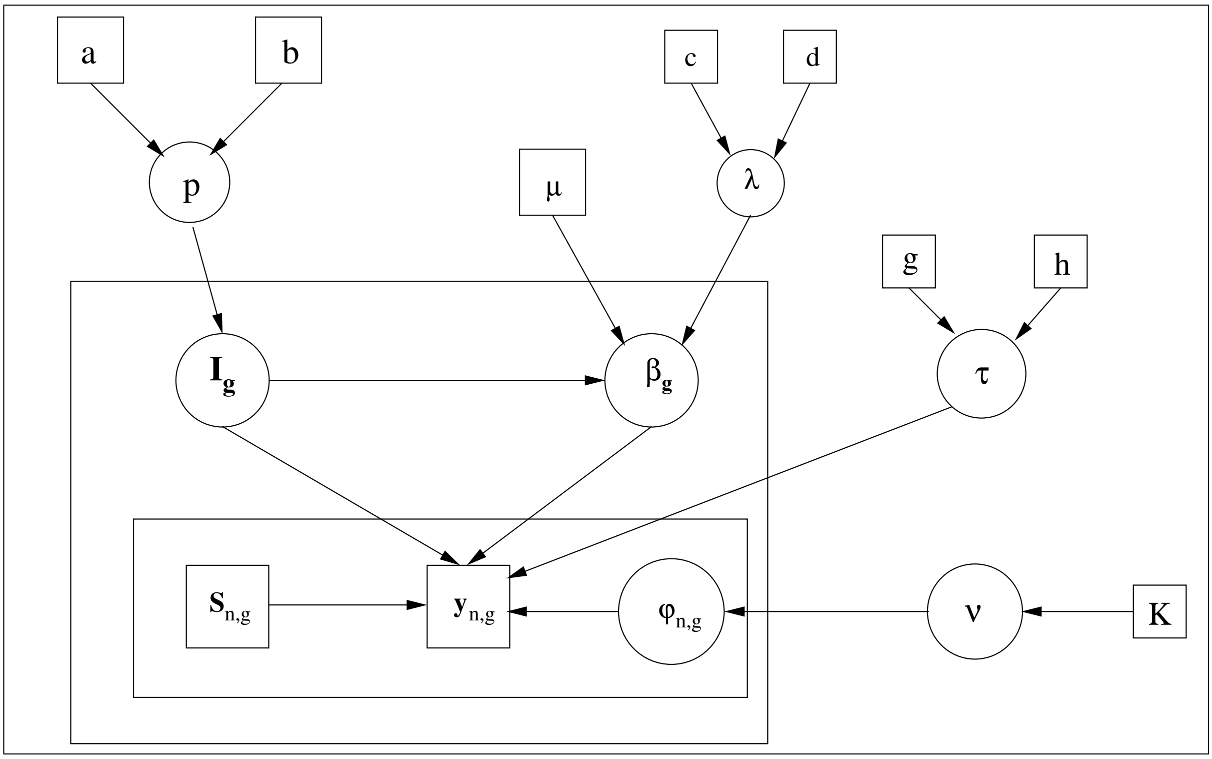


Figure 5.1: Directed Acyclic Graph (DAG) representation of the model; Rectangular frames refer to variables which are fixed during the updates (data, fixed hyper-parameters), whereas variables in circles are updated as parts of the model.

$y_{n,g}$ observations of differential expression, i.e. normalised light intensities

$S_{n,g}$ indicator to which experiment class s observation $y_{n,g}$ belongs

β_g ANOVA coefficient vector for gene g, i. e. the vector of mean expressions

I_g indicator of differential expression

p prior probability of a gene to be differentially expressed

λ prior precision of β_g

τ precision of the residual noise model

$\varphi_{n,g}$ rescaling parameter linking normal and t distribution

ν degrees of freedom of the error model

Figure 5.1 visualises the model by means of a Directed Acyclic Graph. Here, it has to be mentioned that this representation of hierarchical Bayesian models is more common in the machine learning community than in the statistics community. However, this visualisation presents a more efficient and useful way for intuitively capturing the model and will thus be used for all models presented in this thesis.

The student's t noise model of varying degrees of freedom forms an essential part of the model in figure 5.1. This approach also allows us to consider robustness issues regarding outliers or overdispersed data points in the oberservations $y_{n,g}$. The modelling of overdispersion is dominated by the degrees of freedom parameter ν of a t distribution which in our model includes only values up to a maximum value. It is generally known that for large enough values the t distributions will be sufficiently similar to normal distributions. Thus, it is assumed that differing between these distributions does not make any sense after a certain point, therefore we specify a cut-off value ν_{max}. Reaching the maximum value is equivalent to choosing a normal distribution model. However, we do not simply approximate the Gaussian distribution by the $t_{\nu_{max}}$ distribution, instead we employed the exact normal distribution model. For flexibility regarding the choice of the degrees of freedom parameter for the t-distribution, a discrete uniform hyper-prior on the set $\mathfrak{N}$ over the parameter ν is specified:

$$\begin{aligned} \nu &\sim U_{\mathfrak{N}} && (5.8) \\ & \mathfrak{N} := \{x \in \mathbb{R} | 1 \leq x := j \cdot c_{grid} \leq \nu_{max}, j \in \mathbb{N}\} && (5.9) \\ \Leftrightarrow \quad & \mathbb{P}[\nu = k|K] = 1/K, \;\; k \in \mathfrak{N}; \;\; K = |\mathfrak{N}| && (5.10) \end{aligned}$$

The choice of a uniform prior on this finite set also appropriately represents our lack of information regarding the underlying noise model. In order to improve the models readability, we introduced the 'size' K with respect to the counting measure of the set $\mathfrak{N}$ for specifying the uniform distribution.
As discussed in Section 2.1.2, the definition of set $\mathfrak{N}$ (5.9) allows for greater flexibility in choosing the underlying parameter space, thus improving the analysis of robust behaviour. In particular, a large grid size c_{grid} equal to

1 or even 5 allows us to work with clearly distinguishable student's t distributions, whereas refining the grid approximates a continuous setting for ν sufficiently well. The importance of using this discrete model lies in the notion of including the normal model not approximately but exactly, which will be realised by a dimension-changing move. Furthermore, inferring the degrees of freedom parameter introduces the possibility of letting the model itself choose the most suitable error distribution. For these reasons, it is recommended to consider a possibly large number of models.

As discussed in Section 5.1, the biological indicator for differential expression follows a Bernoulli distribution

$$\pi(I_g|p) \quad = \quad p^{I_g}(1-p)^{1-I_g}. \tag{5.11}$$

Here, we applied a conjugate beta prior for the parameter p, which can be interpreted as probability of a gene being differentially expressed a priori

$$\pi(p) \quad = \quad \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}. \tag{5.12}$$

Conditional on the differential expression behaviour, the coefficient vector is determined by a mixture of a multivariate and a univariate underlying distribution, as described above.

As a special case of the general setting, we assume several restrictions for the involved parameters. First, the hyper-parameter μ are fixed, i.e. $\mu_{g,s} = \mu \ \forall g, s$, taking the value of the overall sample mean. The precision of β_g shall be specified by the parameter λ, which by assumption is the common parameter for all prior precision parameters. To remain in the conjugate prior setting, λ follows a Gamma distribution, i.e.

$$\tau_{g,s} \quad := \quad \lambda \ \ \forall g, s \tag{5.13}$$

$$\lambda \quad \sim \quad Ga(c, d). \tag{5.14}$$

This reduces the model parts (5.2) and (5.3) to:

$$\begin{aligned} I_g = 0 \quad \beta_{g,0}|I_g \quad &\sim \quad N_1(\mu, (\lambda)^{-1}) \\ &\qquad \beta_g = [\beta_{g,0}, \dots, \beta_{g,0}]^T \ \in \mathbb{R}^{S \times 1} \\ I_g = 1 \quad \beta_g|I_g \quad &\sim \quad N_S(\mu, (\lambda)^{-1} E_S) \end{aligned} \tag{5.15}$$

$$\beta_g|I_g \sim I_g \cdot N_S(\mu, (\lambda)^{-1} E_S) + (1 - I_g) \cdot N_1(\mu_g, \lambda^{-1}). \tag{5.16}$$

The following table presents an overview over the different model parameters and their respective distributions:

$$\begin{aligned} y_{n,g} \quad &\sim \quad N(x_{n,g}^T \beta_g, (\varphi_{n,g} \tau_\varepsilon)^{-1}) \\ \beta_{g,0}|I_g = 0 \quad &\sim \quad N_1(\mu, (\lambda)^{-1}) \\ \beta_g|I_g = 1 \quad &\sim \quad N_S(\mu, (\lambda)^{-1} E_S) \\ \lambda \quad &\sim \quad Ga(c, d) \\ \tau_\varepsilon|g, h \quad &\sim \quad Ga(g, h) \\ \varphi_{n,g}|\nu \quad &\sim \quad Ga(\frac{\nu}{2}, \frac{\nu}{2}) \\ \nu \quad &\sim \quad U_{\mathfrak{N}} \\ I_g|p \quad &\sim \quad Bin(1, p) \\ p \quad &\sim \quad Be(a, b) \end{aligned}$$

Table 5.1: Overview over Student‘s t model

5.2.1 Likelihood Robustness Considerations

As discussed in section 2.1.2, robustness considerations can aim for different components of a probabilistic model. The main focus of this model is the robustification of the likelihood function of a hierarchical ANOVA model, which provides a certain degree of prior robustness by construction. The standard distribution setting for such a model in the field of bioinformatics would be a Gaussian error distribution (see Ibrahim et al. 2002, hierarchical model Zhao et al. 2008; Bayesian ANOVA for microarrays Ishwaran and Rao 2003).

In contrast to these approaches, several authors, among others Berger 1994, suggested to employ Student's t distributions instead of a Gaussian distribution-based model. In the context of microarrays this approach has been used by Gottardo et al. 2006, who already applied t distributions for performing ANOVA analyses with all kinds of possible expression settings. Analysing all possible settings made their approach very hard if not impossible to apply to general microarray experimental settings. Furthermore, it made the approach little useful for practical bioinformatic analyses and for comparisons of scenarios.

The fact that the student's t distribution has a higher probability mass in its tails makes it a reasonable candidate for models which aim at taking outlying values into account. At the same time the t distribution shares certain properties with the normal distribution, such as symmetry and unimodality, as these properties are important for residuals of a regression model. Thus, the student's t distribution is well applicable for modelling values which behave like Gaussian values except for a higher probability of 'outlyingness'. As we are working in the framework of ANOVA, it is necessary to only take care of outliers in the observations $y_{n,g}$. Here, we have another good reason why this approach is focused mainly on robustification of the likelihood function linked to the observations' behaviour.
In order to show the ansatz of robustification in the framework of Bayesian Robustness studies as performed byBerger 1994 for the purpose of robustification of the likelihood, a class Γ of student's t distribution and normal distributions is defined in the following way:

$$\Gamma = \{\{t_\nu(\mu, \tau^{-1}), \nu \in \mathfrak{N} \setminus \{\nu_{max}\}\}, N(\mu, \tau^{-1})\}. \tag{5.17}$$

As discussed in the previous section, the definition of the set $\mathfrak{N}$ in (5.9) makes this approach very flexible. Choosing only a few values for ν allows us to make clear decisions regarding the data's tendency towards normality, respectively its tendency towards the t distribution, which is the general behaviour of interest for us. A finer grid then makes it possible to have an 'almost' smooth representation of the limited parameter space for the degrees of freedom parameter. This discretisation is one reasonable possibility to take a normal model into account instead of an approximation, which would of course be more similar to the nearest t-distributions than to the normal distribution which the t distribution is actually supposed to approximate. Therefore, a reasonable upper bound for ν is crucial in order to make a clear decision as to when the distribution is sufficiently similar to a normal distribution. From this point onward, there is no longer any need for a robustification w. r. t. outliers. Thus, 'jumping' to a normal distribution model, whenever this upper bound is reached, allows us to accurately represent the importance of using the standard approach in cases where robustification is found to be unnecessary.

The presented model's complex hierarchical structure makes finding an analytic solution virtually impossible, thus the usage of sampling methods will be essential. As the model will be treated using a MCMC algorithm, finding the right balance between reasonable and required robustification and computational practicality is essential. In this case, robustness cannot be studied in the way it has been presented for global robustness, as the variation due to the sampling algorithm will be greater than the variation between the parameters (e.g. β_g) for different model settings (e.g. fixed degrees of freedom for 1 student's t model). Hence, the purpose of the model will rather be to indicate, whether or not there exists a problem in principle with the assumption of normally distributed data. On this assumption further analysis steps would be based. The variable degrees of freedom parameter ν is hereby relevant, as it is supposed to give an answer to this question.

Even though we could choose from a broad class of unimodal distributions for robustification attempts, the class of possible likelihoods is limited to (non-central) t distributions with degrees of freedom varying in a predefined set as well as to normal distributions, in order to have analytically tractable models. This robustification approach mainly focuses on outliers of the observations, as discussed in the previous section. The hierarchical structure of the proposed model ensures that a certain robustness w. r. t. the specification of priors is obtained.

An analysis of robustness regarding the range of the posterior distribution or certain parameter estimates is virtually impossible, as these quantities of interests are determined by Markov Chain Monte Carlo simulation. Hence, more than one run per model has to be performed in order to reduce variation introduced by the simulation method itself. These combined results then represent the estimate of the expected value of model parameters. However, performing all these simulations for all models provided by Γ is neither computationally manageable nor of real practical interest. Therefore, the idea of finite classes, originally devised by Shyamalkumar 2000, is adapted in a way that the hierarchical model itself chooses the 'optimal' model given the data and all other modelling components.

The goal of our approach is to focus on the robustness of the likelihood function of a regression model in the framework of microarray analysis. The need for such considerations arises because of the fact that microarrays often produce widely dispersed data. If we take a look at the commonly used models for determining gene expressions, we find that they are based on Gaussian distribution settings, which provide analytically tractable results (e.g. see Ishwaran and Rao 2003). Baldi and Long 2001 for example use t-tests with appropriate adjustments for the number of tests performed. Other researchers introduced fully Bayesian models based on normal distribution assumptions, compare for example Ibrahim et al. 2002, Zhao et al. 2008 and Gottardo et al. 2003. All these approaches have in common that the high probability of 'extreme' values frequently appearing in microarray data affect the outcomes of the normal distribution model. However, these effects have not been systematically studied before Posekany et al. 2011.

Alternatively, a statistical technique for determining the differential expression of genes, as well as for estimating and controlling error rates by means of non-parametric statistics has been introduced by Tusher et al. 2001. Employing non-parametric methods replaces the restrictive assumptions linked with the normal distribution setting with very general ones, at the cost of losing power of tests. Such a method is robust in the sense of independence of assumptions of underlying parametric distributions. However, this does not represent the kind of robustness we are aiming at in our approach. An additional problem we found regard non-parametric approaches is that sample sizes are often too small to gain reasonable outcomes based on rank statistics. This is why a robust, yet parametric approach is more feasible for microarray data than non-parametrics.

In our approach we wanted to stay close to the parametric model of normal distributions on the one hand, but on the other hand take into account data which deviates from the Gaussian distributions setting, e.g. far outlying data points. However, when working with a linear regression model, we still aim for a symmetric unimodal, ideally parametric distribution as error distribution, which is far more specific than the assumptions of non-parametric methods. Attempts for such models have already been made, mainly focusing on Gaussian mixture distributions (cf. Lewin et al. 2007), rarely on t distributions (cf. Gottardo et al. 2006). In some aspects, our modelling attempt is similar to Gottardo's , cf. Gottardo et al. 2006, yet our approach is more generally applicable. In contrast to the approach by Gottardo et al. 2006, we aim at comparing the model to its normal distribution analogue in order to answer the following questions: Is a student's t model required at all? If it is, how "far away" from a normal distribution is it really in terms of degrees of freedom? In addition, we defined the set of t distributions to include all data points into the model in a more general and flexible way. Firstly, we can differentiate between the various t distributions with clearly different degrees of freedom values, which is useful in principle but might be problematic in other respects. Secondly, we were able to reduce the step size far enough, so that ν can be seen as discretisation of a continuous degrees of freedom parameter, while at the same time we keep the advantages of the discrete setting described above. By introducing different test data sets we will show the advantage of using a smaller step size in addition to a larger one. Additionally, the variable dimension of $\beta_g|I_g$ allows our model to be more generally applicable for various types of microarray experimental settings. Without this property, a systematic study of noise behaviour for a large variety of experimental settings would have been impossible.

5.2.2 Data Collection

When performing in-depth assessment of noise models required for microarray data analysis, two main aspects need to be considered:

- First, it is absolutely vital to make sure that the appropriate noise model's inference remains insensitive to the chosen hyper-parameters. Employing synthetically generated data and dedicated spike-in experiments provide the ideal basis for performing such an analysis, as we know the expected outcome. Being familiar with the true underlying data generating process makes such data particularly useful when assessing the effects of the model's structure and building parts, as well as studying the convergence properties of the Markov chain.

- Second, it is necessary to assess a large collection of microarray data sets covering a wide range of model organisms, experimental settings and measurement platforms, in order to draw fairly generally valid conclusions.

In order to create a set of known scenarios artificial data was drawn from a Gaussian, a t_4 and a t_{10} noise distribution. This data simulated a two-way comparison consisting of 500 hypothetical genes. Each of these genes was assigned to one of five groups, the group membership defining the degree of hypothetical differential expression. The mean structure and fraction of each group's occurrence are listed in Table 5.2, variances hereby took values of 0.1, 1 and 10. In order to generate a situation similar to typical microarray datasets, we simulated 5 replicates per group, resulting in 10 synthetically generated data points per gene. To support our conclusions drawn from syn-

subset i	$\mu_{i,1}$	$\mu_{i,2}$	%
1	-12	12	20
2	-5	5	10
3	-1	1	30
4	-0.5	0.5	20
5	0	0	20

Table 5.2: Depending on sample type which is either 1 or 2, genes from subset i are drawn from distributions with means equal to $\mu_{i,1}$ and $\mu_{i,2}$ respectively. The proportion of genes in subset i is shown in column %.

thetic data regarding the proposed models' validity, we additionally employed our method for the spike-in experiment, cf. Choe et al. 2005, which provides a more realistic test case. For bioinformatical preprocessing we applied MAS 5.0 and vsn (Huber et al. 2003).

To ensure that our findings are not limited by particular choices of data sets, we analysed 14 microarray experiments covering various organisms and measurement platforms. Among other settinga, the data include investigations of plant soil responses, drosophila sleep deprivation, primate dietary comparisons and animal liver metabolism. The data sets used in our assessment are summarised in Table 5.3. These chosen experiments can be identified by their respective Gene Expression Omnibus (GEO) reference number (see Edgar et al. 2002a) and cover various platforms and quantification algorithms (see Table 5.3 column "Prep." for details). All data was normalised by vsn and fed into the algorithm as provided by the owner for the respective data bases.

5.2.3 Considered bioinformatical Normalisation and analysis methods

In bioinformatics, it is commonly acknowledged that results of microarray data analyses can strongly depend on the chosen normalisation method, for example cf. Bolstad et al. 2003. To ensure the correctness of our findings independent of the chosen normalisation approach we repeated the analysis on different subsets of the data in Table 5.3 with different normalisation methods. As not all kinds of data can be used for every normalisation approach, mainly due to availability of raw data, results do not exists for every combination of data set and normalisations. For the comparison of our data sets, we chose the frequently applied methods loess (Yang et al. 2002) and quantile (Bolstad et al. 2003) normalisationdue to their popularity in applied microarray papers.

Recent discoveries shed light on the fact that intensities of highly expressed targets cross-talk to neighbouring probes due to scanner inadequacy (Upton and Harrisson 2010). For this reason, we may expect that Affymetrix probe sets contain outlying measurements in more systematic and frequent ways than previously assumed. The two methods, multi-mgMOS (Liu et al.

2005) and PPLR (Liu et al. 2006), were specifically designed to cope with such problems. To test if such representations could be an alternative to heavy tailed noise models we applied our algorithm to data which were normalised with the multi-mgMOS method as well as the posterior expression estimates, which were obtained by the PPLR method. As both methods assume a Bayesian model with Gaussian noise, we are of course specifically interested in the validity of this assumption and in the question of whether or not these approaches would present a reasonable alternative to other more commonly used normalisation methods.

Due to the frequent occurrence of outliers and other problems with the normal distribution model, non-parametric methods are commonly employed for robust assessment of microarray data, cf. Tusher et al. 2001 or Gao and Song 2005. These approaches are per definitionem not limited by distribution assumptions, thus we compare them against with robust, distribution-based approach, in order to relate it to existing robustification strategies. To ensure comparability we chose ANOVA-like methods, such as the Kruskal-Wallis permutation test (Lee et al. 2005), for which we calculated 10000 permutations, and performed an ANOVA on (aligned) rank-transformed data (Haan et al. 2009). When assessing different noise models for microarray data analysis, it is of great importance to evaluate the impact choosing the wrong noise model has on any biological conclusions drawn from the model compared to the ones based on the most appropriate noise model. The implications of choosing a wrong noise model instead of the most appropriate one can for example be investigated at a higher level of biological abstraction by employing Gene Ontology (GO) term analyses, Ashburner et al. 2000, for the gene lists obtained with the different noise models. For our approach, we applied Fishers exact tests on the GO terms which are related to the selected top ranked genes in order to determine which ones are significantly enhanced in the data. This procedure is the standard approach for GO analysis which is also used in bioinformatical applications such as FatiGO, Al-Shahrour et al. 2004, and DAVID, Dennis et al. 2003. In order to quantify the divergence between the Gaussian noise model and the more appropriate heavy-tailed noise model we compared the absolute amount of differentially expressed genes and significant GO terms. Here, we differentiated between genes and GO terms dependent on the noise model and terms, which are independent of the noise model.

5.2.4 Sensitivity analyses and convergence diagnostics

Hierarchical Bayesian models have already been designed and applied for microarray data analysis, for example cf. work by Lewin et al. 2007, Shahbaba and Neal 2006. The hierarchical Bayesian models' virtue lies in adequately representing the inherent randomness of certain parameters, e. g. the mean gene expression. However, hyper-parameters must be chosen carefully, as any informative choice, such as their weight being high w. r. t. the measurements, will have an impact on the inference results. To our advantage most parameters follow distributions where hyper-parameters have easily understandable meanings. For example in the conjugate Jeffreys prior for the probability of differential expression $p \sim \mathrm{B}eta(a{=}\frac{1}{2},b{=}\frac{1}{2})$ we can interpret a and b as prior observations of (non-)differential expression, weighted with $\frac{1}{2}$. However, other hyper-parameters exert influence more subtly, e. g. the precision parameters λ and τ. For them the improper, yet valid $\mathrm{G}amma(0,0)$ distribution corresponds to the limit case, where the Bayes estimator equals the maximum likelihood estimator. As pointed out in Bernardo and Smith 2000, such theoretically motivated choices are well justified in single variable cases as opposed to a multi-variable model such as the model in Figure 5.1, which deserves further attention.

In contrast to these well-behaved parameters, the precision λ is the only random hyper-parameter which directly influences the mean expressions and thus also the decision about differential expression. Hence, the hyper-parameters c and d in its prior have to be chosen with great care, as the posterior probabilities of differential expression I_g will be quite sensitive to informative settings. In order to investigate the effects of different choices for the hyper-parameters c and d we performed a sensitivity analysis with the artificial data generated according to the description in Section 5.2.2 with precision 1. Varying these two hyper-parameters indicates how much implicitly introduced prior information is tolerated by the model. The graphs in Figure 5.2 illustrate the dependency of the gene-wise posterior probabilities of differential expression on the hyper-parameters c and d as well. As the precision in the Gaussian prior over β_g is modelled hierarchically, the prior variance of $E[(\lambda - \hat{\lambda})^2]_{p(\lambda|c,d)} = \frac{c}{d^2}$ is the determining factor of model sensitivity. By changing the prior variance we are able to assess sensitivity, while keeping the expectation $E[\lambda]_{p(\lambda|c,d)} = \frac{c}{d}$ fixed. Up to a small prior variance of less than 1/500 hyper-parameter values have only moderate influence on the posterior probabilities of differential expression, as can be seen in Figure 5.2. Our improper choice is thus justified, as no local influence on the inference results is introduced.

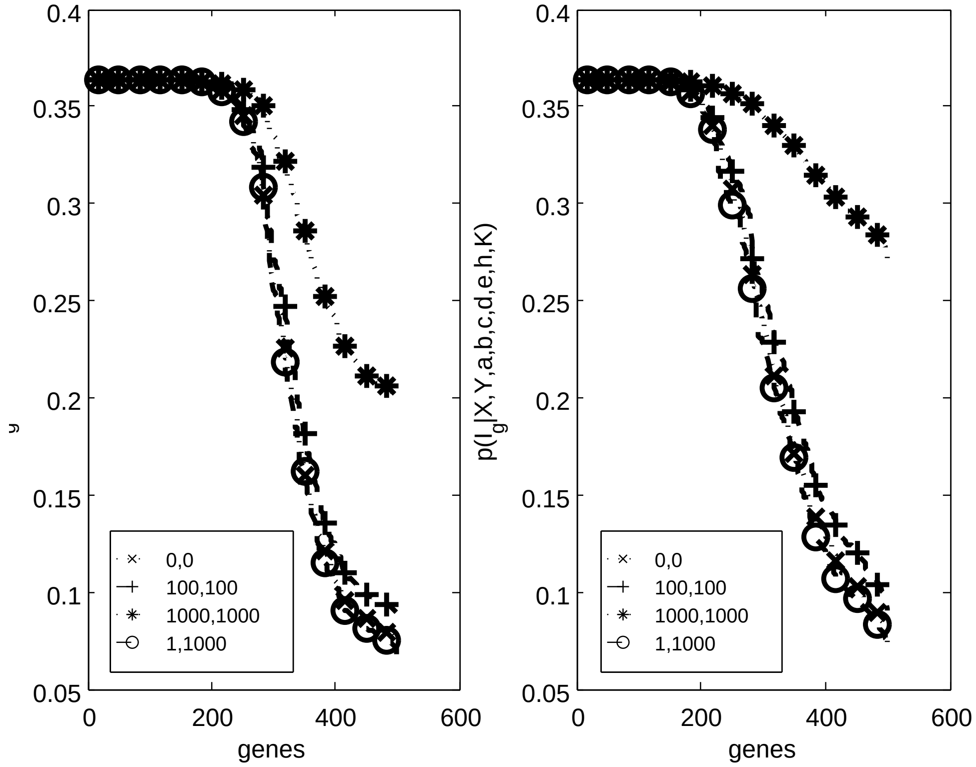
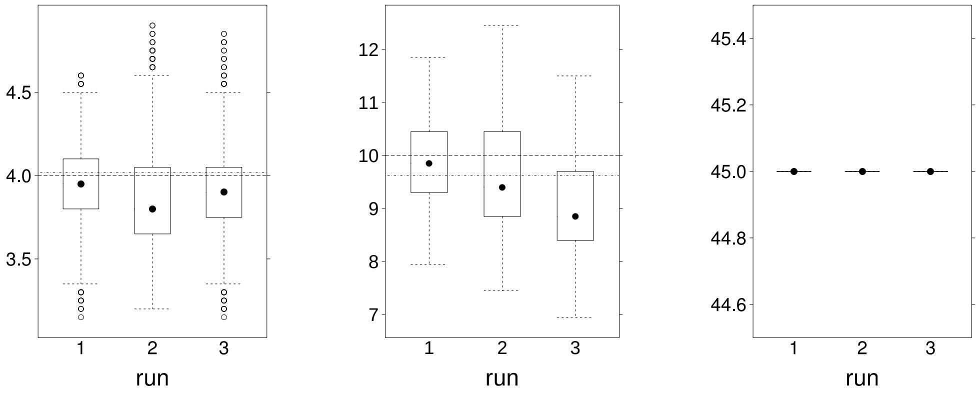


Figure 5.2: Here, the influence of the hyper-parameters c and d in the prior over λ for Gaussian (left) and t_4 (right) distributed data is illustrated by their increasingly different behaviour. The above two graphs show the ranked gene specific posterior probabilities of differential expression for different prior variances of λ.

The cut-off ν_{max} for the degrees of freedom parameters presents another influential hyperparameter in our model. Choosing the upper limit for the degrees of freedom parameter ν, in order to mark the bound between student's t and normal distributions, is critical for clearly distinguishing between



(a) t_4 data (b) t_{10} data (c) normal data

Figure 5.3: The above box plots represent the posterior distribution of the estimated degrees of freedom parameter for a t_4, a t_{10} and a Gaussian data set. The strongly dashed line marks the true degrees of freedom value and the dash-dotted line marks the posterior mean of all three data sets per setting, i. e. 4.21, 10.66, $45 \sim \infty$.

Gaussian and student's t noise models. Our simulations found that student's t densities with degrees of freedom larger than 45 are almost indistinguishable from appropriately parameterised Gaussians, an observations which is consistent with visualisations of the considered distributions. When adding values above 45 to the set ν, the model selection results in large uncertainties in choosing the most proper noise model. Smaller values for this cut-off lead to a misjudging of student's t distributions with relatively large degrees of freedom for Gaussians. Consequently, we decided on $\nu_{max} = 45$ as the threshold for switching to the Gaussian noise model.

5.2.5 Inferring the noise model

The synthetic data sets defined in Section 5.2.2 provided us with a good testing option for checking the most important feature of the algorithm; its ability to correctly determine the underlying error distribution, independent of the data's variance. Another important aspect of MCMC algorithms is the assessment of convergence towards the stationary distribution. To this effect, we applied the R package coda, introduced by Plummer et al. 2006b. Our conclusion was that 11000 draws were a suitable simulation length, whereas the first 500 draws should be considered the burn-in phase. Figure 5.3 visualises the distributions of the samples of ν as box plots. After the burn-in phase, the sampler draws the degrees of freedom parameter ν around the true value, while the variation, which is estimated by the interquartile range, remains less than 2 degrees of freedom in all cases. This way we arrive at a comparatively small range which definitely excludes the incorrect high degrees of freedom as wel as the Gaussian model, in case of student's t data. On the contrary, the MCMC sampler correctly identifies the Gaussian noise model during the burn-in phase for the Gaussian data and stays consequently with the Gaussian model, without ever leaving it again. Because the Gaussian model fits the data exceptionally well, moving back to the more complex 44 degrees of freedom t-distribution model would be highly unlikely. As we have seen the algorithm includes an implicit penalty for moving from the simpler Gaussian model, for which all rescaling parameters $\varphi_{n,g}$ equal 1, to the much more complex t models, for which these rescaling factors have to be inferred additionally. To conclude, the algorithm's ability to identify all error distributions correctly in our test data sets assures us that our method is well-suited for identifying the required robustness level in real microarray data.

Our simulations on artificial data sets also revealed that the flexible adjustment of the grid size c_{grid} during runtime improves mixing and thus as well enhances the convergence properties of the Markov chain. During burn-in phase, the grid size is refined from an initial value in the range of 1 to 5, as proposed in Gottardo et al. 2006, to a smaller value of about 0.05, which remains fixed during the following sampling process. A relatively large grid size of about 1 ensures that the algorithm is able to quickly determine the approximately correct error model. In contrast to a large grid size, reducing the grid size after the first half of burn-in to $c_{grid} \approx 0.05$ improves mixing of the Markov chain without limiting the possibility of the algorithm to reach distant degrees of freedom. Furthermore, defining the set ν flexibly allows inferring the degrees of freedom ν via a discrete random variable J. The introduction of variable J enables the approximation of the continuous *true* degrees of freedom with high accuracy. Moreover, the MCMC sampler with reduced grid size requires less updates in order to reliably infer the degrees of freedom. All in all, varying the grid size results in the improvement of chain mixing and is thus highly recommendable.

As opposed to the artificial data, the ”Golden Spike” experiment by Choe et al. 2005 provides a more realistic test case, as it is a regularised data set, wherein variation mainly originates from laboratory work. When comparing the performance of our algorithm to the ones Choe et al. 2005 analysed (such as Tusher et al. 2001; Baldi and Long 2001), we found that our algorithm's performance, based on the same preprocessing, was at the top end compared to all considered methods. Liu et al. 2006 gained efficiency with their multi-mgMOS normalised data possibly because the normalisation method is specifically tailored for handling outliers in Affymetrix data, cf. Upton and Harrisson 2010. However, when analysing the PPLR model's expression estimates, we found that these are heavier-tailed than the mmgmos normalised input data. Nonetheless, the mmgmos model by Liu et al. 2006 assumes Gaussian distributions, which are inappropriate for inferring heavier-tailed estimates, as summarised by the results in Table 5.6.

Analysing spike-in data produced the compelling result that, for data in which the main source of errors are laboratory processes, a student's t model fits much better than a Gaussian one. When comparing the gene lists determined by our algorithm against genes from the categories in Table 1 of Choe et al. 2005, the student's t model was able to identify 59% to 86% more genes correctly as differentially expressed than the respective Gaussian model.

Org.	GEO ID	Reference	Prep.	N	$\overline{\nu}$	comm.	diff.	comm.	diff.
						genes		GO terms	
A. thal.	GDS3216	(Dinneny et al. 2008)	MAS5.0	12	4.71	1176	150	111	78
A. thal.	GDS3225	(Van Hoewyk et al. 2008)	MAS5.0	4	5.50	832	290	161	21
D. rerio	GDS1404	(Cameron et al. 2005)	PathStat	10	13.58	1776	136	11	14
D. mel.	GDS1686 (I)	(Zimmerman et al. 2006)	RMA	9	3.62	136	174	11	96
H. sap.	CAMDA 08	(Affara et al. 2007)	CLSS4.1	24	4.04	400	304	26	67
H. sap.	GDS1375	(Talantov et al. 2005a)	MAS5.0	70	3.25	6861	3561	160	316
H. sap.	GDS810	(Blalock et al. 2004)	MAS5.0	31	4.37	72	135	9	51
H. sap.	GDS2960	(Yao et al. 2007a)	RGP3.0	101	4.33	318	166	51	2
M. musc.	GDS660	(Small et al. 2005)	MAS5.0	22	10.48	584	126	20	26
M. musc.	GDS3221	(Somel et al. 2008)	RMA	24	4.21	180	119	108	52
M. musc.	GDS3162	(Someya et al. 2008)	MAS5.0	10	4.38	797	446	112	66
M. musc.	GDS1555	(MacLennan et al. 2006)	MAS5.0	8	3.90	131	183	24	110
R. nor.	GDS2946	(Li et al. 2008a)	MAS5.0	15	4.57	146	157	14	306
R. nor.	GDS972	(Jin et al. 2003)	MAS5.0	44	4.98	369	163	94	71
D. mel.	”Spike In”	(Choe et al. 2005)	MAS5.0	6	3.74	401	1748	-	-

Table 5.3: The above overview of the biological data sets describes the organism (Org.), the GEO ID (CAMDA 08 refers to the Endothelial Apoptosis contest datasets of the meeting and ”Spike In” to the ”Golden Spike” experiment), the preprocessing method (Prep.), the overall number of arrays (N), the average degrees of freedom ($\overline{\nu}$), the number of common genes (comm.), the number of genes with noise model depending differential expression assessment (diff.), the number of common GO terms (comm.) and finally the number of noise model dependent GO terms (diff.). The GEO entry GDS1686 (I) refers to the behavioural subset of the data (only the sleep deprived flies). In column prep. we use MAS5.0 to refer to the Affymetrix MAS 5.0 quantisation method, RMA to refer to the “Robust Multi-array Average” method by Irizarry et al. 2003a (both used for Affymetrix arrays), PathStat for referring to the package described in Middleton et al. 2004, CLSS4.1 to refer to the Codelink Software Suite 4.1 and RGP3.0 to refer to Research Genetics’ Pathway software v. 3.0.

5.3 Bioinformatical Results

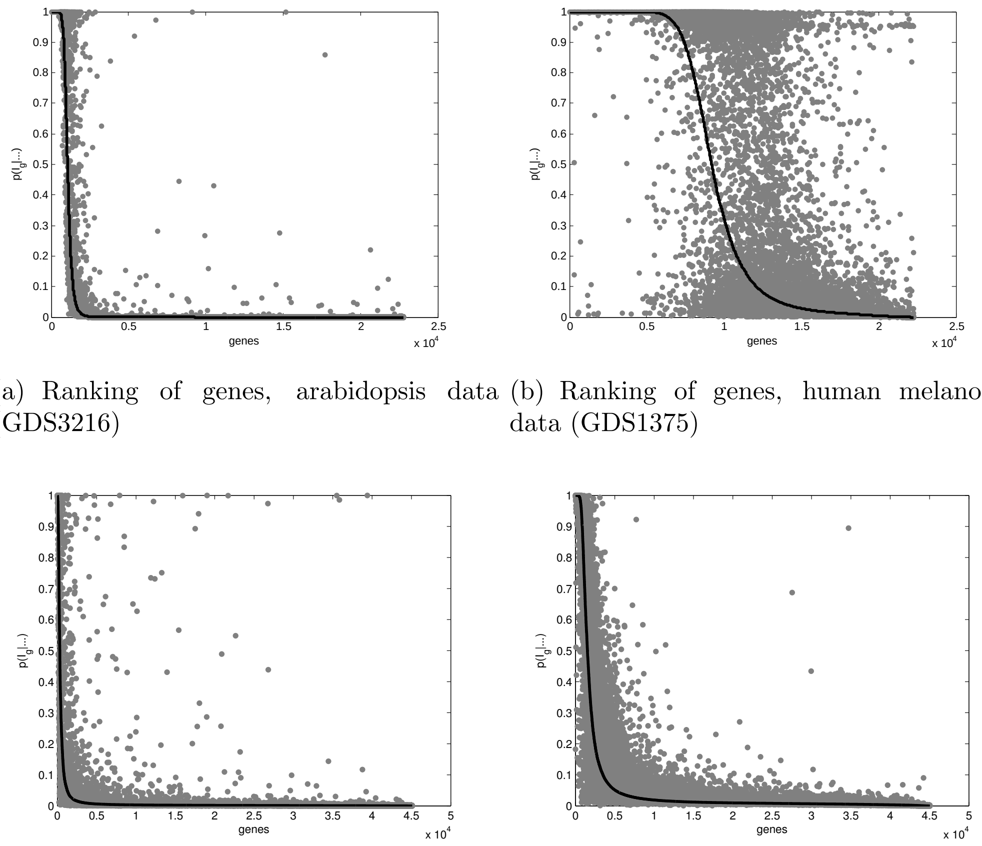
The following section describes our findings for the vsn normalised microarray data sets. In order to highlight the importance of choosing valid noise models for microarray analysis we applied the proposed inference scheme to fourteen microarray data sets, listed in Table 5.3. To obtain a quantitative statement we inferred differentially expressed genes from the Gaussian and the estimated optimal student's t model for every data set. As a result we received two lists of differentially expressed genes, the intersect representing agreement and the symmetric difference representing different biological interpretations induced by an inappropriate noise model.

In addition, we provided Table 5.4 which contains our findings for the alternative normalisations and non-parametric methods. Our evaluation led us to the conclusion that a heavy-tailed Student-t noise model provides a better fit than a Gaussian noise model for every considered data set independent of the normalisation. For most data sets, a student's t distribution with degrees of freedom between 1 and 5 resulted in the highest posterior probability. This clearly indicates the need for robust noise models, which are able to handle outlying data points better than Gaussian model. Thus, we conclude that Gaussian noise models are unsuitable for microarray analysis, even if, according to Novak et al. 2006a, only about 5 to 15 percent of samples are non-normally distributed.

When comparing our parametric approaches against the two considered non-parametric ones we encountered two situations: First, the hierarchical Bayesian model is able to deal with situations in which non-parametric tests have problems with low power, as very few samples or replicates per group are available (cf. Whitley and Ball 2002). This stands out clearly in the four data sets which contain 4 to 24 samples overall, but only 2 or 3 replicates per group. Here, the non-parametric tests were unable to identify any genes as significantly differentially expressed. Second, the robust model shares more top ranked genes of the gene lists with the non-parametric methods than the Gaussian model does, especially in cases of large sample or replicate size. This result indicates that considering to model non-Gaussianity with heavy tails is a reasonable approach in order to describe some of the non-Gaussian behaviour of the data.

By definition, the robust model is generally less sensitive to outlying values, as they are modelled to be closer to the bulk of the data. Models with student's t distributed noise will thus assign lower posterior probabilities of differential expression, if the classification is drawn by one or a few outlying values. In general, fact is that outlying observations increase variance. In cases where outliers additionally lead to a decreased difference between the different average expression values, the Gaussian noise model overlooks differentially expressed genes, which would however be captured by the heavier-tailed noise model. Therefore, we may expect that a wrongly chosen noise model could lead to false positives and false negatives. This expectation is confirmed by the graphs in Figure 5.4, examples which illustrate such noise model dependencies of the posterior probabilities of differential expression for two of the datasets. The above statement remains true independently of the applied normalisation method, as we can see in Figures 5.5, 5.6 and 5.7. Figures 5.4 and 5.5 visualise the graphs for vsn normalised data, Figure 5.6 for the loess normalised data and 5.7 the quantile normalised data.

In the above figures, Figures 5.4 and 5.5, each graph plots the posterior probabilities obtained from a Gaussian or alternatively the most probable student's t noise model against the genes ranked w. r. t. one of the models. When observing the probabilities resulting from the other noise model shown as grey dots, cf. Figures 5.4 and 5.5, we find both false positives and false negatives. On the one hand, we see that several of the genes considered highly differentially expressed by the Gaussian model clearly have a pronouncedly lower posterior probability in the robust model. On the other hand, single genes or whole 'clusters' of genes which have low posterior probability in the Gaussian model are actually highly differentially expressed in the student's t model. The human melanoma (GDS1375) data set in Figure 5.5 (b) presents a good example of a large cluster of such genes, as can be seen at the top right of the graph. As the model inference over degrees of freedom ν clearly favours the robust student's t model we may regard these genes as those which the Gaussian model would have overlooked. To put this in numbers, Table 5.3 lists that the number of genes showing a noise model dependency in the differential expression assessment range from 119 to 3561. This corresponds to approximately one tenth to two times the number of genes which are assessed as differentially expressed, independently of the noise model. As we



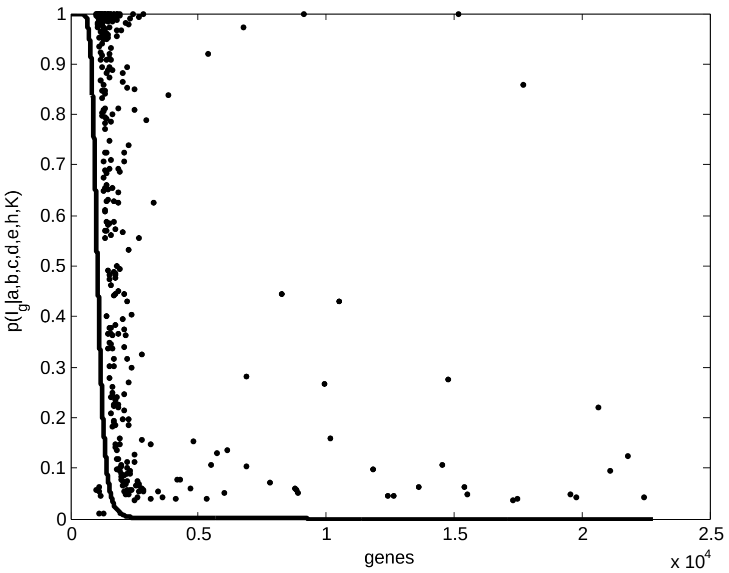
(a) Ranking of genes, arabidopsis data (GDS3216)

(b) Ranking of genes, human melanoma data (GDS1375)

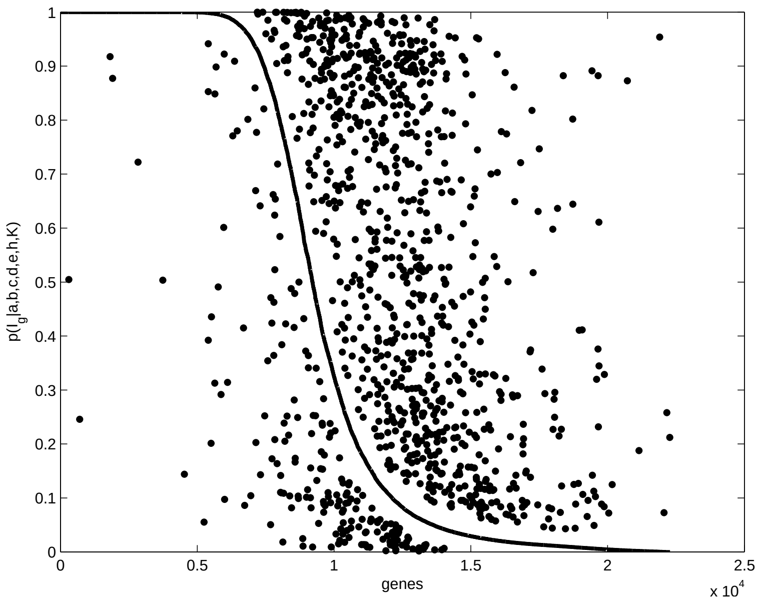
(c) Ranking of genes, mouse glycerol data (GDS1555)

(d) Ranking of genes, mouse cochlea data (GDS3162)

Figure 5.4: The above plots illustrate the noise-model-dependent difference in posterior probability of differential expression. The graph in subplot (a) is ranked by the posterior probability of differential expression obtained with the most probable t-distributed noise model (black line) with corresponding posterior probabilities from a Gaussian noise model shown as grey dots. The graph in subplot (b) is ranked by the posterior probability of differential expression obtained with a Gaussian noise model (black line) with corresponding posterior probabilities of the most probable t-distributed noise model drawn as grey dots.



(a) Ranking of genes, arabidopsis data (GDS3216)



(b) Ranking of genes, human melanoma data (GDS1375)

Figure 5.5: The above two plots present the noise-model-dependent difference in posterior probability of differential expression. The two graphs are ranked w. r. t. the t distribution, as in the previous figure. To make the points with most the prominent differences between the models more visible, we thinned out the points within an interval of 0.1 around the line.

have seen, all results point to the fact that the choice of noise model can be very influential on inferred gene lists.

The biological significance of the differences in gene lists can be further assessed by a Gene Ontology term inference based on a similar strategy as FatiGO (Al-Shahrour et al. 2004). In order to investigate the dependency of such high-level biological inference on the chosen noise model we applied our approach to the gene lists obtained from the Gaussian and most probable student's t model. Comparing the amount of noise-model-dependent and noise-model-independent GO terms, cf. Table 5.3, reveals that between one fifth and 22 *times as many* GO terms differ between the models than the models have common. All in all our results lead to only one conclusion, namely that an unsuitably chosen noise model is likely to have a profound effect on all biological conclusions drawn from a microarray experiment.

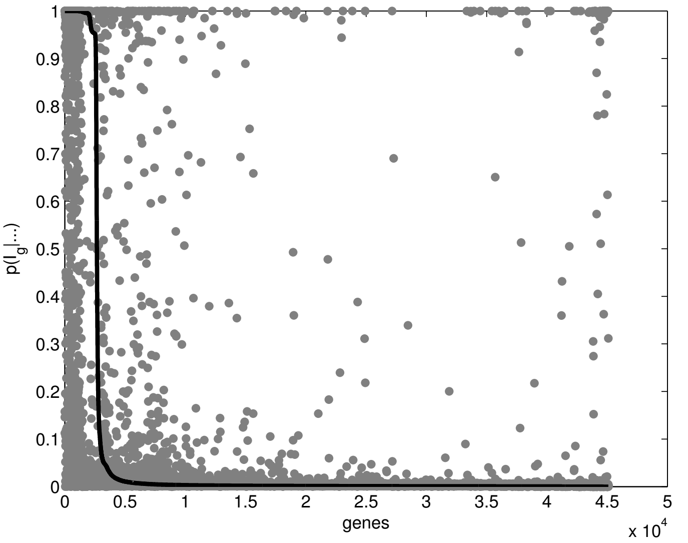
The diverse structure of the experiments, which cover various popular measurement platforms and organisms used for this assessment, suggests that these results will hold in general. Therefore, we conclude that the choice of noise model is likely to have serious implications on inferred gene lists and high level biological conclusions drawn from microarray experiments. Our evaluations, assigning a large posterior probability to student's t distributed noise with rather small degrees of freedom, clearly shows that microarray data analysis requires robust noise models. Consequently, our conclusion is that Gaussian noise, which is often chosen for convenience, is utterly unsuitable for the analysis of microarray data.

5.3.1 Alternative Normalisation

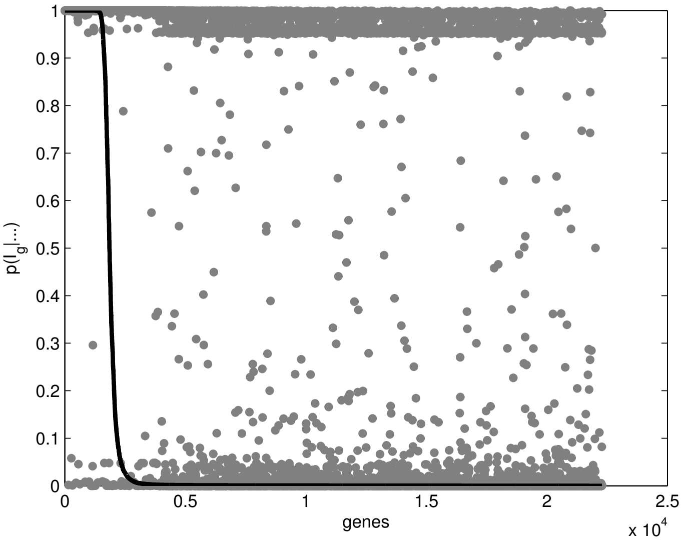
All results presented above are mainly based on vsn normalised data (see Huber et al. 2003). In order to assure that the found effects are not due to this specific normalisation method, we applied rma normalisation (Irizarry et al. 2003b) to those data sets, for which CEL files were available. Furthermore, we selected a subset of data sets to which we applied loess and quantile normalisation as well as Liu's normalisation based on probe-level measurement errors (Liu et al. 2005,Liu et al. 2006). Table 5.4 lists these alternative results in detail. However, the principal findings for vsn normalised data do not change in the least, when applying other normalisation methods to the same data sets: In all cases student's t models with low degrees of freedom are preferred over the Gaussian model. In case of loess and quantile normalised data, in which the variance stabilisation is missing as the overall behaviour of the original data is retained, the degrees of freedom estimate is even lower than for the vsn data. Regarding the human melanoma data (GDS1375), which we discussed above in great detail, a student's t model with about 1.1 degrees of freedom has the highest posterior probability for both loess and quantile normalised data. For loess or quantile normalised data we even found systematic preferences for degrees of freedom of the selected optimal t model, which are much lower than for vsn normalised data. All in all, the posterior over the model showed strong preference for Cauchy-like t distributions with degrees of freedom close to 1, which is the most heavy-tailed distribution available in our finite set Γ, in almost all cases. These observations are consistent with findings by Purdom and Holmes 2005. A possible interpretation for this behaviour could be that loess and quantile normalisation keep the structure of the data's original distribution far better than the variance stabilising normalisation. Thus, the skewed and over-dispersed behaviour of the data is kept, in which the large number of outlying data points is far better explained by very heavy-tailed models. Consequently, the Gaussian distribution cannot provide any reasonable outcomes for data which are that much dominated by outliers. Figures 5.6 and 5.7 picture the results for two of the data sets presented above. Concerning the human melanoma data set, the differences within the posterior probabilities are eye-catching, as a large percentage of genes is classified differently w. r. t. differential expression. The arabidopsis data set represents a more typical case, in which the differences are less prominent when visually comparing the gene lists. However, these differences become more pronounced when performing follow-up analyses, such as Gene Ontology analysis, on these gene lists. Here, the mismatch of the Gaussian distribution becomes highly influential, so that results obtained with Gaussian methods can only be considered unreliable.

5.3.2 Non-parametric methods

Non-parametric methods are generally applied, if the validity of distribution assumptions is unknown or in doubt. Therefore, such approaches are

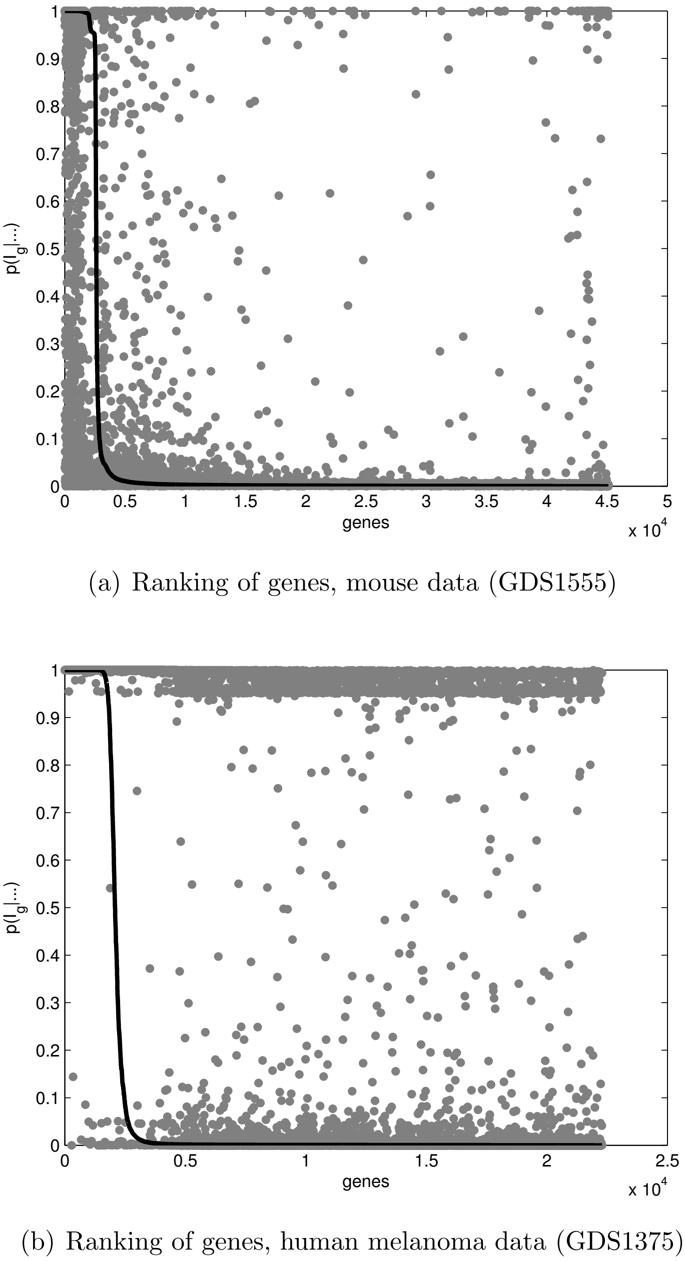


(a) Ranking of genes, mouse data (GDS1555)



(b) Ranking of genes, human melanoma data (GDS1375)

Figure 5.6: The above two plots illustrate the difference in the ranked posterior probability of differential expression for **loess** normalised data. Each graph is ranked separately and the genes on the x axis are ordered w. r. t. decreasing posterior probability in the Gaussian model.



(a) Ranking of genes, mouse data (GDS1555)

(b) Ranking of genes, human melanoma data (GDS1375)

Figure 5.7: The above two plots express the difference in the ranked posterior probability of differential expression for **quantile** normalised data. Each graph is ranked separately and the genes on the x axis are ordered w. r. t. decreasing posterior probability in the Gaussian model.

GEO ID		loess			quantile	
	$\overline{\nu}$	comm.	diff.	$\overline{\nu}$	comm.	diff.
GDS3216	2.02	1273	1272	1.13	1284	1017
GDS3225	1.24	933	1643	1.29	1141	1592
GDS810	1.13	355	860	1.18	487	892
CAMDA 08	1.06	295	1271	1.11	444	1057
GDS1375	1.14	1657	7020	1.15	1863	6881
GDS2960	2.94	268	270	2.85	276	307
GDS1555	1.15	786	2039	1.17	825	1972
GDS972	1.38	545	851	1.4	749	699

Table 5.4: The above table provides us with a subset of data sets for testing alternative normalisations. We calculated the posterior mean degrees of freedom $\overline{\nu}$ and the numbers of common ('comm.') and different ('diff.') genes, which the two methods classified as differentially expressed, with a probability of more than 85%. In all cases, t distributions with small degrees of freedom between 1 and 3 are preferred.

commonly applied for robust assessment of microarray data, see Tusher et al. 2001 or Gao and Song 2005. For analysing microarray data we chose the following non-parametric methods: The Kruskal-Wallis test, the classical non-parametric version of one-way ANOVA, ANOVA on rank transformed and aligned rank-transformed data, as described by De Haan et al. 2009, as well as permutation tests based on (non-)parametric statistics, for example see Lee et al. 2005.

We chose these methods due to their good comparability, as these methods present exactly the non-parametric generalisations of a one-way ANOVA approach. The Kruskal-Wallis test is the non-parametric generalisation of the t test on ranked statistics. However, it works under the assumption of an approximately parametric distribution of its test statistic. In order to avoid this assumption a permutation test can be performed with the Kruskal-Wallis test statistic instead of referring to approximate distributions. We performed such a permutation test using 10000 permutations to estimate the distribution of the test statistic over the data set.

Concerning the robustification of ANOVA De Haan et al. 2009 evaluated several approaches, including rank-transforming the data as well as employing robust mean estimates, for example the truncated mean and the me-

GEO ID	KW perm.		RANOVA (ART)	
	robust	Gaussian	robust	Gaussian
GDS3216	39%	37%	-	-
GDS3225	-	-	-	-
CAMDA 08	-	-	-	-
GDS1375	86%	84%	86%	83%
GDS2960	76%	71%	76%	72%
GDS1555	-	-	-	-
GDS972	35%	35%	36%	35%

Table 5.5: The above table summarises the results of testing non-parametric methods for a subset of data sets. "KW perm" contains the fraction of shared results at a p-value cut-off of 1% for the Kruskal Wallis permutation test with the robust student's t and the Gaussian model, "RANOVA" the fraction of shared results for aligned rank transformed ANOVA with the robust student's t and the Gaussian model, respectively. A dash is used to express that the non-parametric method could classify no gene as differentially expressed with a p-value smaller than 0.01 due to the replicate or sample size being too small.

dian. As we aimed for a non-parametric approach towards ANOVA, we chose to apply the ANOVA on (aligned) rank-transformed data. In the one-way ANOVA setting, such as our, no difference between the different approaches of rank-transforming the data exists. Such differences would only occur for interaction terms, which assume that more than one factor is available and considered in the analysis. The findings for both methods are listed in Table 5.5.

To be able to compare the nonparametric approaches with our parametric approach, we selected all genes with a p-value of differential expression below 1%. In the next step, we chose the same number of top-ranked genes for the robust as for the Gaussian model and calculated the relative amount of shared genes, which were classified as differentially expressed. In cases with large enough sample and replicate sizes, the non-parametric methods generally shared a slightly larger fraction of genes with the robust student's t model than with the Gaussian one. The better agreement of robust methods indicates that non-parametric approaches are to be generally preferred to parametric ones, if the given sample size allows for their application. However, our analysis also revealed a major drawback of non-parametric approaches, which lies in lack of power ,in cases where few samples or replicates are available, a result consistent with the findings by Whitley and Ball 2002. If only 2-3 replicates per group and 4-24 samples overall were provided by the microarray data, the non-parametric methods were unable to identify any significant genes, which is marked by the dashes in Table 5.5). To conclude, our parametric approach outperforms non-parametric approaches on small data sets, whereas non-parametric approaches provide a reasonable alternative only for very large microarray experiments.

5.3.3 Probe-level measurement error

A different approach towards robustification was chosen by Liu et al. 2005 who integrated effects on probe-level into their probabilistic model. This probabilistic normalisation approach was employed in order to estimate the required variables for calculating the probe-level measurement error. In Liu et al. 2006 Gaussian kernels with variance components are fitted, depending on the variation of probe-level measurements. In order to assess the validity of Gaussian model assumptions for this kind of data we applied our algorithm to the posterior mean estimates of the model by Liu et al. 2006. For testing whether or not such representations present an alternative to heavy-tailed noise models, we applied our algorithm to multi-mgMOS normalised data. In addition, we used it on the posterior expression estimates, obtained by the PPLR method, in order to test the model's Gaussian noise assumption. When applying the algorithm to the mmgMOS normalised data our findings were that the over-all noise of the expressions followed a student's t distribution with degrees of freedom between 2 and 3, as listed in Table 5.6. However, results from the PPLR model's expression estimates showed a heavier-tailed distribution than their mmgmos normalised input data, even though the model which infers theses input data assumes Gaussian distributions.

5.3.4 Conclusion

The motivation for our approach lies in the Gaussian distributions' weakness towards outlying and over-dispersed values, which is overcome by non-parametric methods or alternatively by models including heavy-tailed noise distributions, such as the student's t density. Guided by the notion of

	GEO ID		
	GDS3216	GDS810	GDS972
	multi-mgMOS		
$\overline{\nu}$	2.23	3.23	3.67
comm.	815	327	432
diff.	467	354	178
	PPLR		
$\overline{\nu}$	1.17	1.14	1.15
comm.	2504	668	1029
diff.	1045	919	622

Table 5.6: The above table presents the results for the mmgMOS and the PPLR method, separately listing the 3 data sets for which we had CEL files available. The abbreviations 'comm.' and 'diff.' hereby signify the number of common and different genes which are classified as differentially expressed by the robust aa well as the normal model.

Bayesian likelihood robustness, we performed model selection in a way that regards Gaussian noise as well as heavy tailed t-distributions, which allows for a comparison on the level of these different noise models. Once the a posteriori most probable likelihood function i. e. noise model is found, we turned to investigating the biological implications caused by *changing* the noise model. In order to provide conclusions of wide-ranging validity we assessed 14 suitably chosen microarray experiments. Our assessment's outstanding results show that t-distributions with high kurtosis are favoured for every analysed experiment, leading to the conclusion that the choice of error model considerably influences the biological conclusions drawn from the analyses.

Chapter 6

Finite Mixture Models in a nutshell

Before defing a mixture model approach for microarray analysis in Chapter 7, we briefly present the required background knowledge about Bayesian finite mixture modelling. An excellent, application-oriented description of the topic of Bayesian finite mixture modelling approach can be found in Frühwirth-Schnatter 2006, a general description of the (non-Bayesian) theory of finite mixture models in Mclachlan and Peel 2000.

6.1 Bayesian Mixture Models

Finite mixture distributions arise naturally in situations where a random variable Y shows heterogeneity across groups, while being homogeneous within each group. A discrete indicator variable for the groups, Z, will take values in $1, \ldots, K$. The weights of the groups, ω_k, are the relative group sizes, the corresponding parameter for each group is denoted by θ_k. Then, the density of the random variable Y can then be rewritten based on the joint density $p(y, Z|\theta_1, \ldots, \theta_K)$ as

$$p(y|\theta_1, \ldots, \theta_K) = \sum_{Z=1}^{K} p(y, Z) = \sum_{k=1}^{K} \omega_k p(y|\theta_k) \tag{6.1}$$

This definition of mixture distribution allows the straight-forward determination of the mixture's moments

$$\mathbb{E}[f(Y)|Z,\omega,\theta_1,\dots,\theta_K] \;=\; \sum_{k=1}^{K}\omega_k\mathbb{E}[f(Y)|\theta_k], \tag{6.2}$$

if the moments of the component distributions

$$\mathbb{E}[f(Y)|\theta_k] \;=\; \int_{\Omega} f(y)p(y|\theta_k)dy \tag{6.3}$$

exist.

In a Bayesian framework, prior distributions for the group variable Z and the weights ω are required. Let us first assume, we know the true group labelling Z. Then the likelihood $p(Z|\omega)$ is combined with a prior distribution for ω. As $p(Z|\omega)$ is of multinomial structure, the conjugate prior for ω is the Dirichlet distribution $Dir(\alpha_0)$ for ω, cf. 2.1.1.

$$\mathbb{P}[Z=k|\omega]=\omega_k \tag{6.4}$$

$$\omega \sim Dir(\alpha_0) \tag{6.5}$$

Posterior weights ω^* would then be determined as the sum of prior weights, α_0, and the number of observations falling in each of the K groups, $\omega^*_k = \alpha_0+N_k$. If the allocation Z itself is unknown, we require a hierarchical model structure, where for the likelihood function $p(y|Z)$, we assume a multinomial prior over Z with probabilities ω and the conjugate Dirichlet prior for the now hyper-parameter ω. For both cases, the choice of conjugate prior distributions is only one among many. However, it is the most common one, due to the advantage of using Gibbs updates for computational inference which are easy and straight-forward to implement. As an illustration for the conjugate prior setting, we present a small example which we will build on, also when illustrating identifiability, see Section 6.1.1.

Example 5 (Mixture of normals). *We employ the mixture of univariate normal distributions for illustration purposes, not only because it is the classical and most commonly used case of a mixture model, but also because it provides the background of the model designed in chapter 7. The backbone of the model* *is built by normal distributions with mean μ_k and variance σ_k^2 respectively*

$$p(y|\theta = ((\omega_1, \mu_1, \sigma_1^2), \ldots, (\omega_K, \mu_K, \sigma_K^2))) \quad = \quad \sum_{k=1}^{K} \omega_k \phi(y|\mu_k, \sigma_k^2). \tag{6.6}$$

This notation is equivalent to

$$y_i | Z = k \quad \sim \quad N(\mu_k, \sigma_k^2). \tag{6.7}$$

We construct a hierarchical distribution setting, where we place priors on the parameters of the mixture which are commonly chosen as conjugate priors

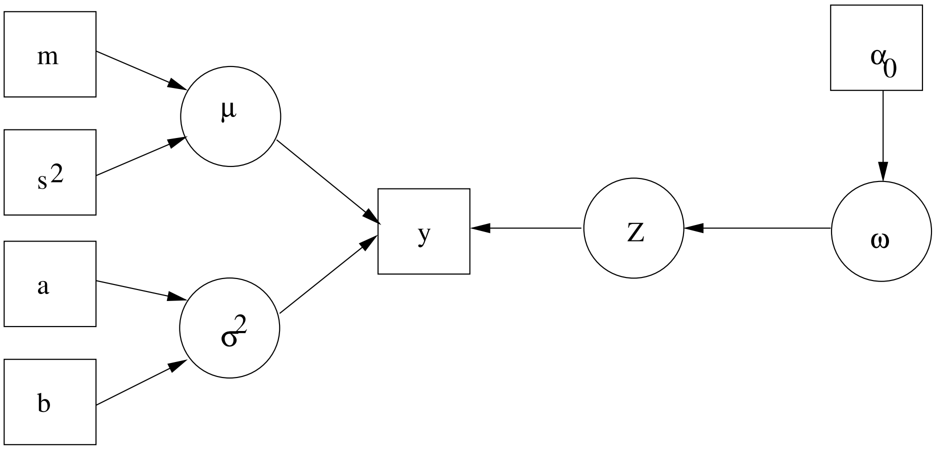


Figure 6.1: The univariate normal mixture model visualised as directed acyclic graph.

$$\begin{aligned} Z &\sim Multinomial(N, \omega) && (6.8) \\ \omega &\sim Dir(\alpha_0) && (6.9) \\ \mu_k &\sim N(\mu_0, s^2) && (6.10) \\ \sigma^2 &\sim Ga^{-1}(a, b) && (6.11) \end{aligned}$$

This is the basic hierarchical normal mixture model, which can be extended in various ways by adding further levels to the hierarchy placing priors on the hyper-parameters or using other types of non-conjugate prior distributions. Figure 6.1 shows the directed acyclic graph representation for this model where we visually separate the two parts of the model:

- *the mixture parameters' part, including Z and ω and*
- *the likelihood's parameters' part, including μ_k and σ_k^2.*

6.1.1 Identifiability

In order to perform sensibly parameter estimation, a model has to be formally identifiable.

Definition 1 (Identifiability)**.** *We define that a parametric distribution on sample space $\mathcal{X}$ with parameter $\theta \in \Theta$ is identifiable, if any two parameters θ and θ' leading to the same probability law on $\mathcal{X}$ are necessarily identical, i. e.*

$$p(x|\theta) = p(x|\theta') \quad \text{for almost all } x \in \mathcal{X} \Rightarrow \theta = \theta' \tag{6.12}$$

Non-identifiability of mixture models can arise from two different sources:

- **Relabelling of the components**
  Any finite mixture distribution is invariant w. r. t. the labelling of the components. Staying with the normal mixture example from above, we can consider a mixture of 2 components where only 2 possible labellings exist. Each defines a distinct parameter $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \omega_1, \omega_2)$ and $\theta' = (\mu_2, \mu_1, \sigma_2^2, \sigma_1^2, \omega_2, \omega_1)$. Thus, the mixture distribution is not identifiable in the sense of the definition above, cf. (6.12). In general, each permutation of the labels $(1, \ldots, K)$ generates the same mixture distribution, despite being defined by distinct parameters. Although this non-identifiability problem is not severe and can be resolved by introducing side-conditions which would allow for identifiability of all distinct parameters which differ in at least one of their components. However, this non-identifiability causes several practical problem which we will discuss in Section 6.2.1 about label switching.

- **Overfitting of components**
  Considering too many components for a finite mixture model also introduces non-identifiability. As shown by Crawford 1994, any mixture with $K-1$ components defines a non-identifiable subset in the K-dimensional parameter space Θ_K of mixture with K components.

Returning to our normal distribution example, we can also add a component of weight 0, defining a $K+1$-dimensional mixture model

$$p(y|\theta_{K+1}) \quad = \quad \sum_{k=1}^{K} \omega_k \phi(y|\mu_k, \sigma_k^2) + 0 \cdot \phi(y|\mu_{K+1}, \sigma_{K+1}^2). \quad (6.13)$$

The new parameter $\theta_{K+1} = (\mu_1, \ldots, \mu_K, \mu_{K+1}, \sigma_1^2, \ldots, \sigma_K^2, \sigma_{K+1}^2, \omega_1, \ldots, \omega_K, \omega_{K+1} = 0)$ is then not identifiable, as μ_{K+1} and σ_{K+1}^2 could take any value and still lead to the same distribution. Alternatively, we can fix these values to those of component j which would then lead to non-identifiability of the weights ω_j and $\omega_{K+1} = 1 - \sum_{k=1}^{K} \omega_k$.

6.2 Computational Inference of mixture models

As in the Example 5 of the finite normal mixture model with fixed number of components, a mixture model can be defined using only conjugate prior distributions, requiring nothing but Gibbs sampling, cf. Frühwirth-Schnatter 2006. Metropolis-Hastings steps can occur for non-conjugate updates leading to a hybrid sampler. However, it is inappropriate to utilise improper priors in the context of mixture models, as they lead to improper posterior distributions. Reusing our example of a mixture of normals, we show how to apply non-conjugate priors for the normal distribution's parameters or mix student's t distributions instead of normal distributions for which the degrees of freedom are unknown.

Example 6 (Mixture of student's t distributions)**.** *Again, we start out from the mixture representation*

$$p(y) = \sum_{Z=1}^{K} p(y, Z) = \sum_{k=1}^{K} \omega_k p(y|\theta_k). \quad (6.14)$$

We now choose student's t distributions as $p(y|\theta_k)$ *and reuse Equation 5.6* *for Chapter 5*

$$X \sim t_\nu(\mu, \sigma^2) \Leftrightarrow \begin{cases} X|\varphi \sim N(\mu, \frac{1}{\varphi}\sigma^2) \\ \varphi \sim Ga(\frac{\nu}{2}, \frac{\nu}{2}) \end{cases} \tag{6.15}$$

Thus, we can consider the mixture model with auxiliary variance rescaling parameters φ

$$\begin{array}{rcr} y|Z = k, \mu_k, \varphi_k, \sigma_k^2 & \sim & N(\mu_k, \frac{1}{\varphi_k}\sigma_k^2) \\ \mu_k|m, s^2 & \sim & N(m, s^2) \\ \sigma_k^2|a, b & \sim & Ga(a, b) \\ \varphi_k|\nu_k & \sim & Ga(\frac{\nu_k}{2}, \frac{\nu_k}{2}) \end{array} \tag{6.16}$$

as a generalisation of the model in Example 5. For fixed degrees of freedom parameters ν_k, we can even stay in the conjugate prior framework and can thus only use Gibbs updates. Only a prior distribution over the degrees of freedom parameters necessarily requires Metropolis-Hastings updates, as no conjugate closed from prior exits for such a case. The Gibbs sampler with auxiliary variable φ leads to samples from the same posterior distribution as an algorithm which uses the proper Metropolis-Hastings step instead, where we use e. g. a standard normal proposal density. See Section 7.2 for detailed updates in such a case.

Even in the Gibbs sampler's case, special challenges arise, when performing computational inference of Bayesian mixture models. One such challenge is the label switching problem.

6.2.1 The Label Switching Problem

The label switching problem is caused by the invariance of the mixture distribution w. r. t. the labelling of its components, cf. Section 6.1.1 where we discussed identifiability issues due to this problem. Figure 6.2 presents the sampled MCMC time series of the mixture weights and degrees of freedom parameters from the mixture of normal and student's t distributions' algorithm in chapter 7 in order to illustrate the label switching problem.

Several approaches exist to deal with this issue which have been categorised in two ways, cf. Crawford 1994 and Stephens 2000b. Firstly, they

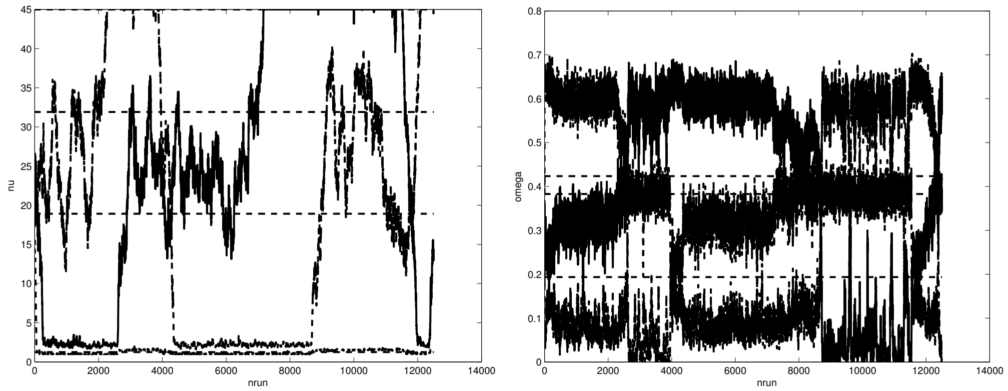


Figure 6.2: Example of fitting a mixture of a normal, t_{10} and t_1 distribution with generated weights $(0.1, 0.4, 0.5)$. Both the degrees of freedom parameter ν_k and the weight parameter ω are affected by label switching, which can be recognised graphically from the crossing of trend lines which exchange their respective trend level, e. g. $\nu = 45 \sim \infty$ and $\nu = 10$

consider the case of a natural ordering of the populations, e. g. when the density means are increasing by assumption. In such a case, there exists only one possible labelling, assuming this ordering as side conditions allows to uniquely identify the components. This provides an optimal solution to the label switching problem, if the choice is theoretically sound and can be reasoned before even fitting a mixture model. However, this might not be the case for many applications of mixture models where the 'true' data generating process remains in the dark. In such cases, we can try discover from projections in the parameter space where it might be possible to identify suitable side conditions which will finally lead to a unique labelling. In addition, recent findings, e. g. by Celeux et al. 2000 and Frühwirth-Schnatter 2001, have revealed that just some arbitrary formal identification constraint does not necessarily lead to unique labelling, presenting counter examples. Frühwirth-Schnatter 2006 recommends the usage of the point process representation, as in Figure 6.2, to identify possibly sensible constraints. Such constraints can be used online while sampling or as as postprocessing device after sampling, "switching back" the labels such that the side conditions are fulfilled for each sample run.

Secondly, Crawford 1994 considers the case that we do not know whether there exists an ordering. Then, a properly introduced identifying function can resolve the problem. An identifying function is defined as a function which maps equivalent values of θ to the same point

$$\begin{aligned} &H : \Theta \to \Theta \\ &p(y|\theta_1) = p(y|\theta_2) \quad \Leftrightarrow \quad H(\theta_1) = \tilde{\theta}_1 = \tilde{\theta}_2 = H(\theta_2) \quad \forall \theta_1, \theta_2 \in \Theta. \end{aligned} \tag{6.17}$$

Then, the distinction between parameters $\tilde{\theta}_1$ and $\tilde{\theta}_2$ implies a distinction between the corresponding distributions $p(y|\tilde{\theta}_1)$ and $p(y|\tilde{\theta}_2)$, thus fulfilling the criterion for identifiabilty, (6.12). The theoretical definition of this function is far simpler than its practical determination. Therefore, researchers typically resolve this issue by applying supervised or unsupervised learning techniques, such as clustering, online relabelling et cetera, cf. Grün and Leisch 2009, Celeux et al. 2000, Frühwirth-Schnatter and Pyne 2011 and Stephens 2000b. Unsupervised clustering is a classical appraoch for undoing label switching in an unconstrained sample. A bridge between the motivation of identifability constraints, which we can consider 'clustering criteria' which could be found manually, and such general unsupervised learning is built by clustering approaches which focus an the point process representation. For all clustering approaches, the MCMC draws in every step are permuted such that a clustering criterion, e. g. k-Means, is fulfilled leading to relabelled samples.

6.2.2 Overfitting

In addition to label switching, another problem arises from the identifiability issues in mixture modelling, cf. 6.1.1. Over-fitting occurs, when more components are fitted than required by the data and issues of separating these components arise. In order to deal with such issues, Frühwirth-Schnatter 2001 has introduced the approach of mode hunting in the posterior point process space. Mode hunting has originally been applied to sample histograms to gain an idea of the number of mixtures to fit, however giving misleading results, as modes in histograms are extremely sensitive to the choice of bin width and number.

The notion of using the mode hunting approach for detecting overfitting is based on the empirical fact that for a large enough number of observations $K!$ dominant modes arise for the mixture likelihood function, if the data are generated by a mixture with K components. However, if the data stem from a mixture with less than K components, i. e. the model is overfitting the data, then less than $K!$ dominant modes are actually present. Searching for these dominant modes, however, can be quite tedious, as the 'clusters' around the modes are blurred with increasing number of overfitting components. Although it makes mode hunting difficult, such blurring is a good indication of possible overfitting.

An alternative approach for resolving overfitting is the method of moments, for which theoretical moments, see (6.2), are compared to sample moments. This method cannot only be used for inference of mixture models with fixed number of components, but also for determining the discrepancy between the data and various such mixture models with different fixed number of components. The true diffficulty then lies in determining when adding another component does not lead to a significant gain in the model fit. In case of doubt, one would rather apply Occam's Razor and stay with the smaller mixture model, avoiding overfitting, cf. Frühwirth-Schnatter 2006.

6.2.3 Transdimensional Methods for Variable numbers of components

When dealing with mixture models, for which the number of components is not fixed a priori, computational inference becomes even more challenging. How to determine the number of components, specifying appropriate priors for Bayesian model selection, can be defined fairly easily. The general framework behind transdimensional approaches is Bayesian model selection. Unlike (6.1), the sampling distribution $p(y|\theta_1, \ldots, \theta_k, \mathcal{M}_k)$ also depends on the model $\mathcal{M}$. In this context models $\mathcal{M}_k$ with different numbers of components k are considered. Given a reasonable prior probability for the model $\mathcal{M}_1, \ldots, \mathcal{M}_K$, $p(\mathcal{M}_k)$, model selection follows Bayes' rule (cf. 2.1, Bernardo and Smith 2000):

$$p(\mathcal{M}_k|y) \quad \sim \quad p(y|\mathcal{M}_k)p(\mathcal{M}_k) \tag{6.18}$$

with the marginal posterior for the model

$$p(y|\mathcal{M}_k) \quad = \quad \int_{\Theta_k} p(y|\mathcal{M}_k, \vartheta_k)p(\vartheta_k|\mathcal{M}_k)d\vartheta_k. \tag{6.19}$$

The computational realisation of such transdimensional methods, however, shows some difficulty, as models with different parameter spaces have to be compared. Here, we briefly discuss two methods of relevance for this thesis which have prominently been applied to deal with this problem, e. g. the reversible jump methodology found one of its first applications in selecting the number of components in a mixture, cf. Richardson and Green 1997, yet it has a broad spectrum of applications also in other fields than mixture modelling, cf. 4.2.3 for the discussion of MCMC updates based on this method.

- **Reversible jump algorithm**
  The purpose of the paper by Richardson and Green 1997 was proposing a novel methodology for determining the number of mixture components in addition to model inference of the mixture distribution. In Chapter 4.2.3 a detailed description of the reversible jump method has already been presented, so this section only focusses on the mixture model aspect of the method. The main issue with the reversible jump approach is the fact that a bijection between superspaces of the model parameters' parameter spaces is highly non-trivial to define. Arbitrarily many possibilities exist, as one may add dimensions at will to the original parameter space in order to construct such a bijection. Even for rather well-known cases, such as multivariate student's t distributions, the definition of a joint update of degrees of freedom, location and scale parameters is not straight forward. Generally, the reversible jump algorithm can move between models of any complexity, i. e. between a model with 3 components and one with 15 components. However, defining the appropriate embedding and bijection is really tedious, if not impossible in some cases. Thus, a special type of reversible jump algorithm has been designed for the purpose of dealing with mixture models with variable number of components: the split and merge algorithm. Instead of moving between arbitrary possible models, only two possible trans-dimensional moves are possible:

  - splitting one of the existing components in two separate components, thus reducing their individual weights
  - merging two separate components into a single one, increasing the

single component's weight.

These are two simpler types of updates, where it is more likely to find an appropriate bijection.

Instead of using split and merge moves, Richardson and Green 1997 also suggested to use birth and death moves. A birth refers to adding a nearly empty component to the mixture which is a theoretically valid move since an empty component may always be added to the mixture model without changing the likelihood function per se. Simply adding a component with weight $\omega_{k'} = 0$ and drawing the respective parameter $\vartheta_{k'}$ from some proposal density however is not a valid move, as there exists no uniquely defined inverse move to form a bijection. The side condition

$$dim(\theta_{K+1}) \quad = \quad dim(\theta_K) + dim(u) \tag{6.20}$$

exists for the parameter u shich relates to the parameters of the added component $K + 1$ in a bijective manner to fulfill the detailed balance condition 4.9 according to the reversible jump construction scheme. Thus, this dimension matching condition is absolutely required for a valid reversible jump move and places clear restrictions on the probability to sample the weights and parameters from. These restrictions led Stephens 1997 to design an alternative approach for such birth and death moves: Stephens' birth-and-death algorithm.

- **Stephens' birth-and-death algorithm**
  To avoid the reversible jump algorithms main down side, the proper specification of a bijection between smartly chosen embedding superspaces of the different dimensional model parameter's spaces and calculate its Jacobian, Stephens 2000a designed an alternative approach, the birth-and-death algorithm. The idea behind this algorithm is to let components "die" and "be born" according to a marked point process, based on the notion that a mixture model might quite abstractly be considered a marked Poisson process in a general space. The birth

rate λ_b of this process stays fixed during sampling and marks when a new component $(\omega_{K+1}, \vartheta_{k+1})$ is added. The probability of death for a component is proportional to the weight of a component, thus, making components with high weights less likely to "die", while those with little weight are prone to be removed from the model. During the birth and death phase, labels are not reassigned, only afterwards the observations are mapped to the new components. In between birth and death phases, an MCMC sampler updates the mixture model with fixed number of components. We present the following scheme as an overview over the birth and death algorithm:

(a) Simulate $(K, \omega, \vartheta_1, \ldots, \vartheta_K)$ by running the birth and death Poisson process for a fixed time T, then set $t = 0$.

- ∗ Determine the current death rate $d(\theta_k)$, proportional to the birth rate λ_b and the ratio of likelihoods and prior model probabilities of the model without and including the component k, and the summed up overall death rate d_t

  $$d(\theta_k) = \frac{p(y|K-1, \theta_{K-1})\lambda_b p(K-1)}{p(y|K, \theta_K) K p(K)}$$

- ∗ Simulate the next arrival time t_{next}

  $$t_{next} = t + \varepsilon/(\lambda_b + d_t) \quad with\, \varepsilon \sim Gamma(1, 1)$$

  as long as $t_{next} < T$ and proceed with step (b) otherwise.

- ∗ Simulate new mixture weights depending on whether a birth or death occurred.
  - · Divide the weight of the dying mixture model among all other mixture models in case of death and
  - · reweight all mixtures' weights such that they sum up to 1 again in case of birth

Table 6.1: Stephens' birth and death algorithm

(a) Adjust the mixture model accordingly, i. e. in case of the death of component k

$$\begin{aligned} \theta_{K-1} &= (\vartheta_1, \dots, \vartheta_{k-1}, \vartheta_{k+1}, \dots, \vartheta_K) \\ \omega &= (\frac{\omega_1}{1-\omega_k}, \dots, \frac{\omega_{k-1}}{1-\omega_k}, \frac{\omega_{k+1}}{1-\omega_k}, \dots, \frac{\omega_K}{1-\omega_k}) \end{aligned}$$

and in case of the birth of component $K+1$

$$\begin{aligned} \theta_{K+1} &= (\vartheta_1, \dots, \vartheta_K, \vartheta_{K+1}) \\ \omega &= (\frac{\omega_1}{1-\omega_{K+1}}, \dots, \frac{\omega_K}{1-\omega_{K+1}}, \omega_{K+1}) \end{aligned}$$

where ϑ_{K+1} is drawn from the prior $p(\vartheta | hyper - parameter)$ and ω_{K+1} simulated from a Beta prior with expectation $\frac{1}{K}$. Set $t = t_{new}$ and return to the beginning of (a).

(b) Update the parameter $(\omega, \vartheta_1, \dots, \vartheta_K)$ and the allocation Z with a fitting MCMC sampler, appropriate for the model with fixed number of components K.

Table 6.2: Stephens' birth and death algorithm

The transdimensional MCMC has the same purpose as Bayesian model selection of the most appropriate mixture model, based on the marginal likelihoods of the different dimensional models. Theoretically, both approaches are equivalent, yet their computational realisation is not. Again both methods are limited by computing power and ressources w. r. t. the maximal number of components to consider and MCMC draws to compute. However, when the number of considered possible components becomes large, i. e. more than 10 to 20 possibilities, transdimensional methods become advantageous and likely the only possible approach, as calculating the marginal likelihood for every single model is virtually impossible, even in the age of parallel computing.

Chapter 7

Extending Bayesian Mixture models

7.1 Model structure

The following hierarchical Bayesian mixture models generalise mixtures of Student's t and Gaussian distributions, respectively. Unlike previously considered models (see e.g. Frühwirth-Schnatter 2006) we mix Student's t and normally distributed components simultaneously. With this approach we allow the model to collapse to the well-known cases of only Student's t or Gaussian distributions, but also cover all cases in between. Instead of density estimation, we focus on identifying systematic overdispersed behaviour. This is particularly relevant in settings of joint inference as found in genomics, e. g. microarray studies. In this case, the mixture model allows for an interpretation of relative amounts of genes with specific noisy behaviour unlike the model selection type of approach, described in chapter 5. While mere density estimation can felxibly and relatively easily handled by mixtures of normals, a severe problem for such an approach would lie in determining the proper number of components in order to stay identifiable. On the one hand, a mixture model which only differs w. r. t. variance parameters is a lot more prone to overfitting than one differing in more than just one parameter. As several normal components would be required to approximate the tail behaviour of a single student's t component, this is a grave disadvantage. On the other hand, fitting only mixtures of student's t distributions, which are more appropriate for microarray data according to Hardin and Wilson 2009 and Posekany et al. 2011, contains the problem of sensitivity to the choice of prior over the degrees of freedom. This prior is extremely sensitive to the cut-off of the maximal degree of freedom when the student's t distributions is already close to the normal distribution. Inference over the degrees of freedom parameter would then also be almost uninterpretable as one has to decide when to consider all degrees of freedom above a threshold as approximately normal. In contrast to approximating a Gaussian distribution with student's t components with high degree of freedom, only the necessary model parameters are inferred. Jumping to the actual normal model instead of considering a whole set of t models as approximation is a logical alternative. Utilising student's t components merely if required by the data structure, inference becomes more simple compared to a mixture of only t distributions. This also has a direct influence on the performance of the corresponding sampling algorithm and the required computational resources, if implemented accordingly.

For illustration purposes, we build up the intended model for microarray data gradually first defing submodels. The first such model (7.1) fits a univariate mixture of Gaussian and t distribution where the components can differ regarding their mean μ_j, precision λ_j and degrees of freedom ν_j, if they are t-distributions. All Gaussians are considered as limiting cases of student's t distributions, where the degrees of freedom are infinite, $\nu_j = \infty$. In the following, we assume an i.i.d. sample $(x_1, \ldots, x_n)$ of size N that is drawn from a mixture of J distributions.

$$x_i \sim \sum_{j=1}^{J} \omega_j f(\mu_j, \lambda_j, \nu_j) \tag{7.1}$$

To make the corresponding likelihood function easy to handle, we used a commonly applied latent variable approach (see Frühwirth-Schnatter 2006 and the explanations in chapter 6). Thereby, we introduced auxiliary variables Z_i that label each observation x_i as being drawn from component j. The probability for label Z_i to be j is equal to ω_j. Thus, the number of observations N_j with label $Z_i = j$ follow a multinomial distribution.

$$\begin{array}{lcll} x_i|Z_i = j & \sim & f(\mu_j, \lambda_j, \nu_j) & \\ \mathbb{P}[Z_i = j] & = & \omega_j & \\ (N_1, \ldots, N_J) & \sim & MN_{N;(\omega_1,\ldots,\omega_J)} & N = \sum_{j=1}^{J} N_j \end{array} \tag{7.2}$$

The component-probabilities $\omega = (\omega_1, \ldots, \omega_J)$ follow a Dirichlet distribution. These assumptions lead to a conjugate prior setting.

We used the relation between the student's t distribution and the normal and Gamma distribution to base the model on these two easier to handle distributions, which also allow us to use conjugate priors. This relation is for example described in (Bernardo and Smith 2000 and chapter 5). The random variable x follows a normal distribution with mean μ and precision, i.e. inverse of the variance, $(\lambda\varphi)$ where the original precision is rescaled by a parameter φ. The rescaling parameter originates from a Gamma distribution with shape and rate parameter $\nu/2$.

$$\begin{array}{lcl} x & \sim & N(\mu, \frac{1}{\lambda\varphi}) \\ \varphi|\nu & \sim & Ga(\frac{\nu}{2}, \frac{\nu}{2}) \end{array} \tag{7.3}$$

Then the marginal distribution of the random variable x is a non-central Student's t distribution of ν degrees of freedom.

$$x \sim t_\nu(\mu, (\lambda)^{-1}) \tag{7.4}$$

Another way of interpreting this relation is that the student's t distribution is a scale mixture of normal distributions where the continuous weighting function is the Gamma distribution. Based on this relation it becomes clearer that a finitie mixture of Gaussian distributions will approximate the t distribution with a certain error depending on the degrees of freedom and the number of fitting components with this error becoming larger for small degrees of freedom, cf. Gelman et al. 2003.

Our distribution model thus contains rescaling factors $\varphi_{i,j}$ which depend on the component j via the degrees of freedom ν_j. In case of Gaussian components all φ_i equal 1 and thus need not be inferred. This is an advantage over estimating high degrees of freedom student's t components.

The model of Gauss-t-mixtures

$$\begin{aligned}
x_i &\sim \sum_{j=1}^{J} \omega_j f(\mu_j, \lambda_j, \nu_j) \\
x_i|Z_i = j &\sim N(\mu_j, \frac{1}{\lambda_j \varphi_{i,j}}) \\
\varphi_{i,j}|\nu_j &\sim Ga(\frac{\nu_j}{2}, \frac{\nu_j}{2}) \\
\mu_j &\sim N(m, \tau^{-1}) \\
\lambda_j &\sim Ga(a, b) \\
b &\sim Ga(c, d) \\
\nu_j &\sim U_{[1;\nu_{max}]} \\
(N_1, \ldots, N_J) &\sim MN_{N;(\omega_1, \ldots, \omega_J)} \quad N = \sum_{j=1}^{J} N_j \\
\omega = (\omega_1, \ldots, \omega_J) &\sim Dir(\delta, \ldots, \delta)
\end{aligned}$$

The second model (7.5) is designed to fit a mixture of distributions as residuals of a linear model, e. g. an ANOVA model. In this case, the mean is estimated by the linear model and thus independent of the mixture component. Therefore, the mean of all components becomes 0, $\mu_j = 0 \quad \forall j$, and the components can only be differentiated by their precisions λ_j and degrees of freedom ν_j.

$$y_i = \beta^T x_i + \varepsilon_i \quad \varepsilon_i = \sum_{j=1}^{J} \omega_j f(0, \lambda_j, \nu_j) \tag{7.5}$$

The residual model allows us to robustify the linear model w. r. t. incoherent underlying noise behaviour. Such behaviour can originate from errors during the experiment or measurement process. Under these circumstances, recognising the respective data points and dealing with them by mapping them to a very noisy error component is sufficient. However, in some cases over-dispersed data points represent samples from a sub-process of the one generating the data. In this case we need to reasonably include information about this behaviour into our model. The suggested mixture model can handle both cases. However, the weight or down-weighing of certain information for estimation or decision making has to be introduced in a reasonable loss function. A possible issue of the approach is identifiability of the two noise-related parameters, describing variance and tail weight in dependence of ony another. Without the presence of a non-noise related variable, such as the component wise mean, introducing identifiability constraints are hard to introduce, while overfitting looms over the data analyst as soon as more than only 2 or 3 interpretable components are fitted. Thus, this approach stays reasonable only for mixutres with few components. In our case, the 'tail weight', expressed by the degrees of freedom parameter is the parameter of inference which we can also use as constraint for discerning the components.

In addition to the mixture model considerations above, our third model is a special case of the residual model, which has been specifically designed to analyse microarray data and answer the biologically relevant question of differential gene expression.

ANOVA model of Gauss-t-mixtures for microarrays

$$\begin{aligned}
y_{n,g} &\sim \sum_{j=1}^{J} \omega_j f(X\beta_g, \lambda_j, \nu_j) \\
y_{n,g}|Z_i = j &\sim N(\mu_j, \frac{1}{\lambda_j \varphi_{n,g,j}}) \\
\varphi_{n,g,j}|\nu_j &\sim Ga(\frac{\nu_j}{2}, \frac{\nu_j}{2}) \\
\beta_g|I_g &\sim I_g N_S(\mu, \tau^{-1}E_S) + (1 - I_g)N_S(\mu, \tau^{-1} \cdot \mathcal{I}_S) \\
I_g|p &\sim Bin(1, p) \\
p &\sim Be(a, b) \\
\tau &\sim Ga(c, d) \\
\lambda_j &\sim Ga(e, h) \\
\nu_j &\sim U_{\{[1,\nu_{max}],\infty\}} \\
(N_1, \ldots, N_J) &\sim MN_{N;(\omega_1,\ldots,\omega_J)} \quad N = \sum_{j=1}^{J} N_j \\
\omega = (\omega_1, \ldots, \omega_J) &\sim Dir(\delta, \ldots, \delta)
\end{aligned}$$

In this model $y_{n,g}$ represents the preprocessed observations, I_g the indicator of differential expression, β_g the vector of mean expressions of all considered biological states. E_S is the S-dimensional unit matrix and $\mathcal{I}_S$ an S-dimensional vector of ones, as described in chapter 5. Fig 7.1 shows the structure of the hierarchical model represented by a directed acyclic graph. The goal is to model the noise as simply as possible while simultaneously including noisy observations and dealing with complicated noise behaviour typical for this type of data. Microarrays are a typical example for a setting in which the over-dispersed data points may contain information about underlying measurement errors in addition to measurement errors. A one-way ANOVA model compares the mean expressions of each gene between the S systems considered in the experiment. Similar to Posekany et al. 2011, cf. chapter 5, we model the mean expression of a gene β_g as a one-dimensional random variable which is identical for each biological system, if the gene is not

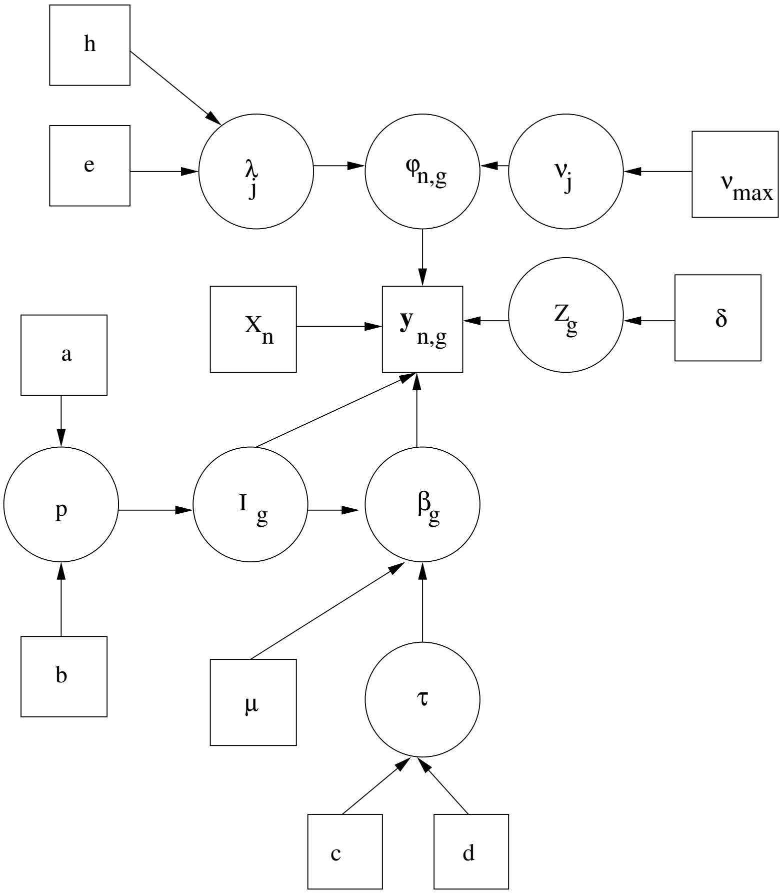


Figure 7.1: Directed acyclic graph representation of the normal-t mixture model for microarray data. The parameters have the following meaning and interpretation within the model or biological system

$y_{n,g}$ . . . observations for gene g and sample n
x_n . . . dummy vector encoding which biological system sample n belongs to
β_g . . . vector of the mean expressions of the compared biological systems
I_g . . . indicator of differential expression
τ . . . precision of the expression within the biological systems
Z_g . . . factor of the mixture component j for each gene g
$\phi_{n,g,j}$ . . . rescaling factor for the observation $y_{n,g}$ given mixture component j
ν_j . . . degrees of freedom of the noise distribution of mixture component j
λ_j . . . precision of the distribution of mixture component j

differentially expressed. This is the null hypothesis of our ANOVA model. In case of differential expression the mean expression follows a multivariate normal distribution where the biological states are considered to be uncorrelated. The mixture of normal and student's distribution is used for the residuals of the linear ANOVA model. Thus, each gene can show a different type of scattering around its mean expression, depending on its weights of the mixture components. As the mean expression is typically dependent on the experiment, the mean vector and scalar μ is estimated from the data as the overall mean expression of all arrays of a given biological state and the whole experiment respectively. Considering the typically small sample sizes in microarray experiments, pooling the information over thousands of genes for estimating the noise components is advantageous.

In addition, we have extended the above models by allowing the number of mixture components to vary for trial pruposes and performed model selection by inferring each model's posterior distribution. Recently, Frühwirth-Schnatter and Pyne 2011 have discussed several reasonable approaches for selecting the number of components in finite mixture models. A valid approach is performing model selection with respect to the number of components. Thus, before inferring the above model with fixed number of components, we extended our models by varying the number of mixture components to perform "a priori" model selection. To this respect, we employed Stephens' (Stephens 2000a) method of variable components estimation. As we use his algorithm to infer the number of components, we applied the same prior setting that he has chosen in his thesis (Stephens 1997) and paper (Stephens 2000a). The number of components are a priori assumed to follow a truncated Poisson distribution.

$$J \sim P_\gamma \mathcal{I}(\{1, \ldots, 100\}) \tag{7.6}$$

The choice of parameter γ is very influential, but interpretable. Selecting a small value of about 2 or 3 will favour mixtures with few components, whereas a large γ will lead to complex mixture with many components. Depending on the setting and aim of an analysis, both choices might be reasonable. When fitting models for high-dimensional problems with thousands of regressors, as in the case of microarrays, a mixture with more than 10 components can be considered. However, fitting smaller and simpler models is generally preferable.

7.2 Markov Chain Monte Carlo Algorithm for the mixture model

The above models are too complex to analytically gain a tractable solution of the parameters' joint posterior. Thus, for inferring our models, we implemented hybrid Markov Chain Monte Carlo (MCMC) samplers, which contain Gibbs, Metropolis-Hastings and Reversible Jump steps and in addition birth and death steps for adding and removing components for the variable number of components algorithm. Robert and Casella (Robert and Casella 1999) among others give a detailed description of the MCMC methods, applied in our sampler. In case of a variable number of components, we applied Stephens' algorithm (Stephens 2000a, Stephens 1997) for getting an indication of the optimal number of components. Furthermore, we showed that convergence concerning the variable number of components algorithm is a complex matter. Even for a converging algorithm, it is hard to perform inference based on the results, as only for the same number of model parameters calculation of posterior is possible. Thus, the number of samples is usually too small or sampling takes too long. Therefore, we estimated the number of components from this algorithm, but inferred the parameters for fixed number of components, based on the a posteriori most likely number of components.

Following Frühwirth-Schnatter (Frühwirth-Schnatter 2006), we used Gibbs updates for all mixture related steps of the algorithm. However, all parameters of interest in the microarray model require non-Gibbs updates. For changing between the different dimensional spaces in the ANOVA model which represent the two hypotheses of interest, we require a reversible jump move. Green (Green 1995) and Richardson and Green (Richardson and Green 1997) have described and applied this type of update in detail. The normal distribution can be considered a special case of the student's t one, which is reached by collapsing the model to estimation of only the mean and precision λ_j with $\nu_j = \infty$ and $\varphi_{n,g} = 1$. Thus, for jumping between the higher dimensional student's t model and the normal model we apply a reversible jump update as well, as described before in Posekany et al. (Posekany et al. 2011).

We use a Metropolis-Hastings step to update the degrees of freedom parameters ν_j, as no conjugate prior is available for these parameters of student's t distributions. Unlike our previous work (Posekany et al. 2011), we have chosen a continuous prior distribution for the ν_j. Thus, we do not approximate, but can directly estimate a continuous posterior distribution. The jump to the truly Gaussian model however does not occur at the maximum value, but in an interval of length 0.5 with limit ν_{max}, $[\nu_{max} - 0.5; \nu_{max}]$. We require an interval to actually perform a jump to the normal distribution with positive probability. The width of 0.5 is chosen dua to the implementation of the uniform prior updates, recommended e. g. by Frühwirth-Schnatter 2006. Instead of sampling uniformly from the whole considered interval $[1, \nu_{max}]$, we draw from a uniform prior on a symmetric interval of length 1 around the current value. This way, the mixing of the sampler is improved, as the proposed value is more likely to be accepted, if it is not too far away form the current one. This implementation bears some similarity to the discrete updata of the degreees of freedom, described in chapter 5 and Posekany 2009.

Additionally, we employ the methodology of partially collapsed Gibbs sampling (cf. Dyk and Park 2008, Park and Dyk 2009) for improving the efficiency of our updates.

In the case of variable numbers of components, we employ Stephens' algorithm (cf. Stephens 2000a), based on random birth and death of components. Compared to reversible jump, the algorithm has the advantage that it is not necessary to construct a bijection between higher dimensional spaces, in which the original ones are embedded, and calculate its inverse. Thus, it can be applied in much more general cases than the reversible jump algorithm. It has been specifically designed for mixtures of t distributions with fixed degrees of freedom but variable number of components where no feasible reversible jump update has been constructed yet.

In the following sections we will describe the updates for the different parameters in details. Based on the theory derived by Besag (Besag et al. 1995) who described the generalisation of MCMC updates of Metropolis-Hastings or Reversible jump type based on full conditionals, we use the notation "$\theta|\Theta_{-\theta}$"

for conditioning θ under all other parameters in the model. Due to the conditional independence property of the DAG representation, this includes only the 'parents' and 'children' in the graph, the parameters directly above and below the given parameter in the hierarchical model which confer to the likelihood and the prior respectively, see Section 2.1.1 on DAGs and hierarchical models.

7.2.1 The noise model parameters

Update ν and $\varphi_{n,g,j}$

Following Frühwirth-Schnatter Frühwirth-Schnatter 2006, we chose a uniform prior on the interval $[1, \nu_{max} = 45]$. For switching between the student's t models, we take a simple Metropolis-Hastings step. To increase the rate of accptance and mixing of the chains, we have limited the proposal of ν to an interval of length 2 around the current degrees of freedom value. At the limits, this probability is capped, resulting in the proposal function

$$\nu^{(new)} = \begin{cases} \nu^{(old)} - u & \nu^{(old)} = \nu_{max} \\ \nu^{(old)} + u & \nu^{(old)} = 1 \\ \max(1, \min(45, \nu^{(old)} + (2 * u - 1))) & \text{else} \end{cases} \tag{7.7}$$

for a uniform random number u.

In addition to student's t distributions, the commonly used Gaussian model was taken into account. To switch between the two types of likelihood a reversible jump step was introduced jumping between the t-model equal to a normal-gamma-model with the auxiliary variables $\varphi_{n,g}$, and the Gaussian model, the limiting case of student's t model, when the auxiliary variables all equal one and the degrees of freedom becomes infinite. In our case, a jump to normal distribution model is proposed when landing at the maximum value ν_{max} in the above model.

Acceptance probability

The acceptance probability for the Metropolis-Hastings move between two different student's t models is

$$\begin{aligned} A &= \frac{\prod_{n,g;Z_g=j} \frac{\Gamma\left((\nu_j^{(n)}+1)/2\right)}{\Gamma\left(\nu_j^{(n)}/2\right)} \nu_j^{1/2} \left(1+\frac{(y_{n,g}-x_n^T\beta_g)^2\lambda_j}{\nu_j^{(n)}}\right)^{-\frac{\nu_j^{(n)}+1}{2}}}{\prod_{n,g;Z_g=j} \frac{\Gamma\left((\nu_j^{(o)}+1)/2\right)}{\Gamma\left(\nu_j^{(o)}/2\right)} \nu_j^{1/2} \left(1+\frac{(y_{n,g}-x_n^T\beta_g)^2\lambda_j}{\nu_j^{(o)}}\right)^{-\frac{\nu_j^{(o)}+1}{2}}} \\ &\quad \cdot \frac{p(\nu^{(n)})}{p(\nu^{(o)})}\cdot\frac{p(\nu^{(o)}|\nu^{(n)})}{p(\nu^{(n)}|\nu^{(o)})} \end{aligned}$$

As we have a uniform prior, $p(\nu^{(o)}) = p(\nu^{(n)})$. Our proposal setting for $p(\nu^{(o)}|\nu^{(n)})$ is given in equation (7.7) and equals the probability of proposing a given $\nu^{(n)}$, when you currently have value $\nu^{(o)}$. For easier reading we denote the logarithmised acceptance probability.

$$\begin{aligned} \log A &= N_j \left[\log\left(\Gamma\left(\frac{\nu_j^{(n)}+1}{2}\right)\right) - \log\left(\Gamma\left(\frac{\nu_j^{(n)}}{2}\right)\right) + \log\left(\Gamma\left(\frac{\nu_j^{(o)}}{2}\right)\right)\right. \\ &\quad \left. - \log\left(\Gamma\left(\frac{\nu_j^{(o)}+1}{2}\right)\right) + \tfrac{1}{2}\left(\log\left(\nu^{(o)}\right) - \log\left(\nu^{(n)}\right)\right)\right] \\ &\quad -\frac{\nu_j^{(n)}+1}{2}\sum_{n,g;Z_g=j}\log\left(1+\frac{(y_{n,g}-x_n^T\beta_g)^2\lambda_j}{\nu_j^{(n)}}\right) \\ &\quad +\frac{\nu_j^{(o)}+1}{2}\sum_{n,g;Z_g=j}\log\left(1+\frac{(y_{n,g}-x_n^T\beta_g)^2\lambda_j}{\nu_j^{(o)}}\right) \\ &\quad +\log p(\nu^{(o)}|\nu^{(n)}) - \log p(\nu^{(n)}|\nu^{(o)}) \\ N_j &= |\{g : Z_g = j\}| \end{aligned}$$

In the special case of the reversible jump step from t distribution to normal distribution similar formulae apply, where the asymmetry of the two different dimensional spaces has to be taken into account.

- Moving from t to Gauss:

$$\begin{aligned}\log A &= -\frac{1}{2}\lambda_j \sum_{n,g;Z_{n,g}=j} (y_{n,g} - x_n^T \beta_g)^2 + g^{*(o)} \sum_{n,g;Z_g=j} \log {h^*_{n,g}}^{(o)} \\ &\quad + N_j \left(\Gamma\left(\frac{\nu_j^{(o)}}{2}\right) - \log \Gamma(g^{*(o)}) - \frac{\nu_j^{(o)}}{2} \log \frac{\nu_j^{(o)}}{2} \right) \\ &\quad + \log p(\to \text{normal model}(\nu = 45)) - \log p(\to t_{\nu_j^{(o)}})\end{aligned}$$

- Moving from Gauss to t:

$$\begin{aligned}\log A &= \frac{1}{2}\lambda_j \sum_{n,g;Z_{n,g}=j} (y_{n,g} - x_n^T \beta_g)^2 - g^{*(n)} \sum_{n,g;Z_g=j} \log {h^*_{n,g}}^{(n)} \\ &\quad - N_j \left(\Gamma\left(\frac{\nu_j^{(n)}}{2}\right) - \log \Gamma(g^{*(n)}) - \frac{\nu_j^{(n)}}{2} \log \frac{\nu_j^{(n)}}{2} \right) \\ &\quad + \log p(\to t_{\nu_j^{(n)}}) - \log p(\to \text{normal model}(\nu = 45))\end{aligned}$$

Here, the parameters g^* and $h^*_{n,g}$ come from the update of $\varphi_{n,g,j}$. The auxiliary variables $\varphi_{n,g,j}$ follow the following Gamma distribution which determines the shape and scale parameters used in the aceptance rate:

$$\begin{aligned}\varphi_{n,g,j} | \Theta_{-\varphi_{n,g,j}} &\sim Ga(g^*_j, h^*_{n,g,j}) \\ g^* &= \frac{\nu_j + 1}{2} \\ h^*_{n,g} &= \frac{1}{2}(\nu_j + \lambda_j (y_{n,g} - x_n^T \beta_g)^2)\end{aligned}$$

7.2.2 Update λ_j

The error model's precision lambda is updated by a Gamma distribution in the following way:

$$\begin{aligned} \lambda_j|\Theta_{-\lambda_j} &\sim Ga(c+\frac{N\cdot N_j}{2}, d+\frac{1}{2}\sum_{n,g;Z_g=j}\varphi_{n,g}(y_{n,g}-x_n^T\beta_g)^2) \\ N_j &= |\{g:Z_g=j\}| \end{aligned}$$

7.2.3 Update allocations Z

Auxiliary allocation variables have been introduced to infer the mixture model in a straightforward way. As each gene is supposed to follow a different underlying noise structure, the allocation is gene dependant, remaining identical over the samples. This update is partially collapsed and uses the student's t likelihood functions instead of the normal appraoximation with rescaling factor phi.

$$\mathbb{P}[Z|\Theta_{-Z}] = \begin{cases} \omega_j\lambda_j^{\frac{N}{2}}\nu^{-\frac{N}{2}}(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})})^N\prod_n(1+\frac{\lambda_j(y_{n,g}-x_n^T\beta_g)^2}{\nu_j})^{-\frac{\nu+1}{2}} & t_{\nu_j} \\ \omega_j\frac{\lambda_j}{2}^{\frac{N}{2}}\prod_n e^{-\frac{1}{2}\lambda_j(y_{n,g}-x_n^T\beta_g)^2} & \text{normal} \end{cases}$$

7.2.4 Update ω

The weights of the components foloow are Dirichlet distribution:

$$\begin{aligned} \omega|\Theta_{-\omega} &\sim Dir(\delta+N_1,\ldots,\delta+N_J) \\ N_j &= |\{g:Z_g=j\}| \end{aligned}$$

7.2.5 Updating the ANOVA model parameters

An updating move for the parameter p is made by drawing p from the updated Beta distribution

$$p|\Theta_{-p} \sim Be(a+i_1, b+(G-i_1))$$

where I is the vector of all I_g and $i_1 = |\{g : I_g = 1\}|$, i.e. the number of genes, which are differentially expressed.

The hyperparameter τ will be updated in the following way:

$$\begin{aligned}
\tau &\sim Ga(f^*, h^*) \\
& f^* = f + \frac{G - i_1 + i_1 * S}{2} \\
& h^* = h + \frac{1}{2}[\sum_{g;I_g=0} (\beta_{g,0} - \mu)^2 + \sum_{g;I_g=1} (\beta_g - \mu)^T (\beta_g - \mu)]
\end{aligned}$$

Updating β_g and I_g

(WD) update β_g conditional on all other variables

case $I_g = 0$

$$\beta_{g,0} | I_g = 0, Z_g = j, \Theta_{-(\beta_g, I_g, Z_g)} \quad \sim \quad N_1(\mu_0^*, (\tau_0^*)^{-1})$$

$$\begin{aligned} \mu_0^* &= \frac{\lambda_j \sum_{n=1}^N \varphi_{n,g} y_{n,g} + \tau\mu}{\tau_0^*} \\ \tau_0^* &= (\lambda_j \sum_{n=1}^N \varphi_{n,g} + \tau) \end{aligned}$$

case $I_g = 1$

$$\beta_g | I_g = 1, Z_g = j, \Theta_{-(\beta_g, I_g, Z_g)} \quad \sim \quad N_S(\mu^*, (\tau^*)^{-1})$$

$$\begin{aligned} \mu^* &= (\tau^*)^{-1}(\tau\mu + \lambda_j X D_{\varphi,g} Y_g^T) \\ \tau^* &= \tau I_S + \lambda_j X D_{\varphi,g} X^T = diag(\tau_1^*, \ldots, \tau_S^*) \\ & \quad with\ \tau_s^* = (\lambda_j \sum_{i=1}^N \varphi_{n,g}^{(s)} + \tau) \end{aligned}$$

The covariance matrix τ is always diagonal, because the matrix X is the design matrix of the ANOVA model, indicating the group s a sample n belongs to, as only in the diagonal two non-zero entries are multiplied with one another.

(RJ) case $I_g = 1 \to I_g = 0$: the proposal for $\beta_{g,0}$ is as above

$$\beta_{g,0}|I_g = 0, Z_g = j, \Theta_{-(\beta_g, I_g, Z_g)} \sim N_1(\mu^*, (\varphi\tau^*)^{-1})$$

$$\begin{aligned} \mu^* &= \frac{\lambda_j \sum_{n=1}^{N} \varphi_{n,g} y_{n,g} + \tau\mu}{\tau^*} \\ \tau^* &= (\lambda_j \sum_{n=1}^{N} \varphi_{n,g} + \tau) \end{aligned}$$

The auxiliary variable A denotes the acceptance probability and uses $\mu_0^*, \tau_0^*, \mu^*, \tau^*$ as defined above for the within-dimension update

$$\begin{aligned} A &= \frac{\prod_n p(y_{n,g} - \beta_{g,0}|I_g = 0, \ldots)}{\prod_n p(y_{n,g} - x_{n,g}^T \beta_g | I_g = 1, \ldots)} \\ &\quad \frac{p(\beta_{g,0}|\mu_{g,0}, \tau_{g,0}, I_g = 0) p(I_g = 0)}{p(\beta_g|\mu_g, T_g, I_g = 1) p(I_g = 1)} \\ &\quad \frac{p(\beta_g | I_g = 1, \ldots) p(I_g = 1)}{p(\beta_{g,0}|I_g = 0, \ldots) p(I_g = 0)} \\ &= \tau^{-\frac{S-1}{2}} \frac{1-p}{p} \frac{\sqrt{|\tau^*|}}{\sqrt{\tau}} e^{-\frac{1}{2}(\tau\mu^2 + \mu^{*T}\tau^*\mu^* - S\tau\mu^2 - \tau_0^*\mu_0^*)} \end{aligned}$$

This results in an acceptance probability of α=min$\{1,A\}$.

case $I_g = 0 \to I_g = 1$: the proposal for β_g

$$\beta_g|I_g = 1, Z_g = j, \Theta_{-(\beta_g, I_g, Z_g)} \sim N_S(\mu^*, (\tau^*)^{-1})$$

$$\begin{aligned} \mu^* &= (\tau^*)^{-1}(\tau\mu + \lambda_j X D_{\varphi,g} Y_g^T) \\ \tau^* &= diag(\tau_1^*, \ldots, \tau_S^*); \tau_s^* = (\lambda_j \sum_{i=1}^{N} \varphi_{n,g}^{(s)} + \tau) \end{aligned}$$

leads to an acceptance probability of α=min$\{1,A^{-1}\}$

7.2.6 Theoretical consideration of Convergence

The samplers based on the models presented in this thesis are hybrid samplers, containing some Gibbs-steps, Metropolis-Hastings updates and mixture kernels of Gibbs and reversible jump updates. Partially collapsed sampling is employed in addition, resulting in more efficient MH steps, when Gibbs updates would have been a straightforward to implement alternative. The composed transition kernel for the model presented in chapter 5 would look like this:

$$\begin{aligned}\mathcal{K}_{hybrid} &= \left(\tfrac{1}{K}\mathcal{K}_{RJ}^{(\nu),(\phi_{n,g})} + \tfrac{K-1}{K}\mathcal{K}_{MH}^{(\nu),(\phi_{n,g})}\right) \circ \mathcal{K}_{G}^{(\tau)} \circ \mathcal{K}_{G}^{(p)} \\ &\circ\left(\mathcal{K}_{G}^{(\lambda)} \circ \left(0.5 \cdot \mathcal{K}_{G}^{(I_g),(\beta_g)} + 0.5 \cdot \mathcal{K}_{RJ}^{(I_g),(\beta_g)}\right)\right)\end{aligned}$$

The kernel for the current sampler is extended by the mixture model resulting in

$$\begin{aligned}\mathcal{K}_{hybrid;mixture} &= \left(\mathcal{K}_{RJ}^{(\nu_j),(\phi_{n,g,j})}\right) \circ \mathcal{K}_{G}^{(\lambda_j)} \circ \mathcal{K}_{G}^{(Z_g)} \circ \mathcal{K}_{G}^{(p)} \\ &\circ\left(\mathcal{K}_{G}^{(\tau)} \circ \left(0.5 \cdot \mathcal{K}_{G}^{(I_g),(\beta_g)} + 0.5 \cdot \mathcal{K}_{RJ}^{(I_g),(\beta_g)}\right)\right)\end{aligned}$$

Briefly, we discuss the individual kernels of the algorithm to discern the properties of such a sampler, using this particular one as a practical example. Here, we describe and argue the properties, which we theoretically deduced according to the theory of section (4.1), for each kernel and their compositions, based on works by Roberts and Sahu 2001, Roberts and Rosenthal 2006, Roberts and Rosenthal 1998a and Roberts and Rosenthal 1998b. However, one has to keep in mind that due to the complexity of the MCMC sampler theoretical results are only partially feasible and some mean of "extrapolation from what is rigorously proven", as Roberts and Rosenthal put it, is required.
As Besag et al. 1995 argued, we can view these kernels as full conditional kernels, thus, defining a proper sampler. The reversible jump steps form special extensions of Metropolis-Hastings steps in such a way that their line of argumentation includes them into the full conditional scheme.

Firstly, we take a closer look at the Metropolis-Hastings type kernels, including the Gibbs steps, since each of them can be viewed as special type Metropolis-Hastings step. The first property to consider is the *support* of the proposal and target / posterior distribution. An important condition is that the support of the proposal distribution contains the support of the posterior. According to Theorem 12, this is sufficient for the sampler to fulfil the *detailed balance condition* and for the posterior to be the stationary distribution of the chain. Naturally, all Gibbs steps fulfil this property, since distributions of the same family and structure with identical support are involved. In case of the Metropolis-Hastings step, which in this sampler update the degrees of freedom ν and the rescaling factors $\varphi_{n,g}$, we define the support of the proposal distribution as the set $R^+ \times \mathfrak{N}$, the same as the support of the desired posterior, cf. Posekany 2009 and Section 5. All reversible jump steps fulfil the detailed balance condition by construction. Since each kernel fulfils the detailed balance condition, we can deduce the existence of an *invariant density*, as both properties are equivalent.

Secondly, we want to assure that we do not only have the chance to explore the whole support, but will also do so 'often enough' in all places to properly simulate the posterior distribution. One such required property is *irreducibility*. According to MCMC theory, presented above, this is fulfiled for every kernel, if the proposal distributions are positive on the support of ξ. Again, this is naturally the case for all considered distribution, since all proposal distributions are either continuous probability distributions with the same support as the posterior or positive by construction, as in case of ν. However, Roberts and Rosenthal 1998a argue that the irreducibility of the hybrid sampler cannot simply be deduced from irreducibility of each component. Due to the complexity which makes an analytical treatment of this question impossible, only detailed simulation studies allow us to answer this question approximately. Irreducibility combined with the existence of invariant density implies *positivity* of the chain per definitionem. Theorem 6 states that positivity implies recurrence. Moreover, it implies *Harris recurrence* for all Metropolis-Hastings type kernels, and additionally provides the condition required for the Convergence theorem (13). Regarding the cycle of kernels, irreducibility and aperiodicity are inherited from a single component with this property. Recurrence follows from recurrence of each of the components, as the term is defined by reaching a set infinitely often, which is composed of subsets with only this property. The existence of a dominating measure for all of the invariant densities is a requirement for the existence of an invariant density of the hybrid kernel. But as the kernels do not have a common invariant distribution little can be stated in general about this invariant density of the hybrid kernel. A generic prove for the existence of such an invariant measure and convergence is however beyond the scope of this thesis. Given the complexity of the sampler at hand, proving the conditions for theorem 14 analytically is virtually impossible. Approximate results can again only be obtained based on the samples and corresponding convergence diagnostic tools, cf. 4.3 and 5.2.4.

7.2.7 . . . and what we can do in practice

After having discussed some theoretical aspects of convergence for this hybrid sampler in the spirit of chapter 4, we focus now on practical aspects of convergence and applications for determining convergence of MCMC samplers, which are in the focus of Cowles and Carlin (Cowles and Carlin 1996) and Robert (Robert and Casella 2009) and have been summarised in 4.3. In his thesis (Stephens 1997) and paper (Stephens 2000a), Stephens deals with theoretical and practical convergence aspects of his algorithm, as do Cappe et al. (Cappe et al. 2003). Both find that convergence is tricky to show and cannot be proven in general. Matching this, we observe only weak convergence of the inference results for the variable component mixture after 15000 samples per chain in 5 parallel chains. Longer chains are not feasible for microarray data, but for smaller amounts of data it is an option to collect enough samples for performing inference directly with this algorithm. Therefore, we consider the found posterior of the number of mixture components only as an indication for which fixed numbers of components to take into account. Our inference of the noise and differential expression is always based on models with a fixed number of components.

Following advice by Robert Robert and Casella 2009, we have assessed convergence of the algorithm with a fixed number of components using the CODA package (Plummer et al. 2006a) in R (R Development Core Team 2011). Thus, we could determine that a burn-in length of 2500 draws and

10000 samples from each of 5 parallel chains, resulting in 50000 draws, are sufficient for posterior inference of the considered parameters.

7.2.8 Simulated data

In order to test the performance, convergence and properties of the algorithm we generated test data sets. For the ANOVA scenario we created two groups of samples, cf. 7.1.

Table 7.1: Structure of the test data sets; In all cases we have 2 groups of ANOVA results: one with means $\mu_1 = \mu_2 = 0$ the other with $\mu_1 = -5, \mu_2 = 5$. The variance lies between 1 and 25 and overall we simulated 500 artificial "genes", 20% of which are differentially expressed.

J	ν_1	%	ν_2	%
2	∞	80 | 50	1	20 | 50
2	∞	80 | 50	4	20 | 50
2	∞	80 | 50	10	20 | 50

J	ν_1	%	ν_2	%	ν_3	%
3	∞	80 | 60 | 10	7	10 | 20 | 40	∞	10 | 20 | 50
3	∞	80 | 60 | 10	10	10 | 20 | 40	4	10 | 20 | 50
3	∞	80 | 60 | 10	4	10 | 20 | 40	1	10 | 20 | 50
3	∞	80 | 60 | 10	7	10 | 20 | 40	1	10 | 20 | 50
3	∞	80 | 60 | 10	10	10 | 20 | 40	1	10 | 20 | 50

One group has the two means $\mu_1 = \mu_2 = 0$ and the other $\mu_1 = -5, \mu_2 = 5$. The residuals are simulated as mixtures of 2 or 3 different distributions (t_1, t_4, t_7, t_{10} and normal distribution) with different weights. The settings with high weights on the normal component (0.8 and 0.6) represent our originally intended setting, while the setting with only 10% of normally distributed data corresponds to the scenario observed for microarray data.

7.2.9 Sensitivity analysis and robustness

With our approach we aim for robustness in several ways. On the one hand, we want to perform robust modelling of noisy and over-dispersed data with appropriate noise models. On the other hand, we formulate our model to be robust regarding our choice of priors by creating a hierarchical model with hyper-priors on the model parameters which is a foremost goal for any computational Bayesian inference. Although the respective hyper-parameters have only little information, they still draw the whole posterior analysis, if not chosen well. By performing a sensitivity analysis we wish to estimate the influence of this choice on our inference results. Here, we discuss this check for the the precision or scaling parameters which are most crucial parameters influencing our analysis.

For the sensitivity analysis we specifically focused on two sets of influential hyper-parameters, the gamma hyper-parameters (c,d) and (e,h). The parameters c and d determine the prior of the rate parameter in the gamma distribution which models the precision of the mixture components. Varying these parameters will influence the recognition of the proper noise model in the components and can lead as far as favouring Gaussian models only or just very extreme models, i. e. the Gaussian and t_1 model. Our tests have shown that such a trend exists. Here, we observed local robustness of our weakly informative choice of parameters around $c = 0.1$, $d = 0.1$. Figure 7.2 visualises this behaviour for two of the test data sets.

Increasing these parameters to $c = d = 10$, $c = d = 100$ and $c = d = 1000$ shows a constant trend where only the two extreme models, Gauss and t_1, are favoured. The noise component estimation results in a Gaussian and a t_1 component and additionally up to one or more other components, which vary a lot, but have a weight close to 0 and are thus results of over-fitting. The larger the influence of these prior parameters becomes, the more the differential expression analysis of the microarray model is affected. For $c = d = 1000$ the posterior estimator of differential expression is almost unable to discern between differential expression and non-differential expression, i. e. between the 2 ANOVA hypotheses.

The second influential set of parameters contains the prior parameters for the rate of the underlying precision of the differential expression, i. e. the group mean of the ANOVA. A wrong choice of these parameters will have a more direct effect on the test for differential expression. Again we show that our prior choice of $e = h = 0.1$ is locally robust. As described above, we observe that the posterior distribution of the indicator of the ANOVA hypothesis, i. e. differential expression, collapses starting from $c = d = 100$.

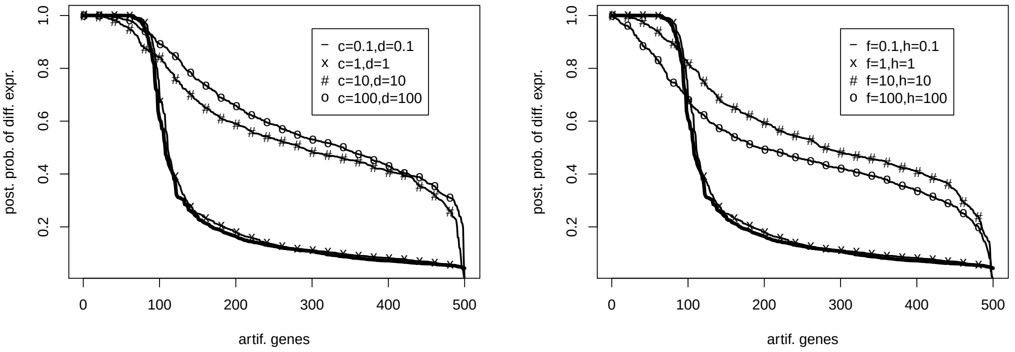


Figure 7.2: Plots of the sensitivity analysis for changing the parameter set (c,d). For the test data set with 10% Gaussian, 40% t_4 and 50% t_{10} data (to the left) and 10% Gaussian, 40% t_{10} and 50% t_1 data (to the right) the plot visualises how the differential expression assessment is affected by changing the prior towards a more informative setting. We can see that our choice $c = d = 0.1$ is locally robust.

Modelling the residuals, equal to the difference of the observations and the respective group means, with student's t distributions we gain robustness against outlying values.

7.2.10 Measuring Non-Gaussianity

We introduce a measure for non-Gaussianity in our analysis for two reasons. First, we want to deal with the label-switching problem in an efficient and straightforward way. This problem is introduced, because the Bayesian mixture model is a priori not identified and we did not force any identifiability constraints on our model, cf. chapter 6. Second, we aim for measuring the "non-Gaussianity" of each gene to sum up its noise behaviour in a reasonable way.

Yau and Holmes 2011 recently published a loss scheme for hierarchical Bayesian nonparametric mixture models in order to determine the relevance of variables. Contrary to their scheme, we wish to remain with the parametric setting, but penalise straying from normality. Thus, we base our measure of non-Gaussianity on the concept of peakedness, which discerns the heavy-tailed student's t distribution from the Gaussians.

As the 4^{th} moment does not exist for the most interesting and "non-Gaussian" distributions with degrees of freedom $\nu \leq 4$, we cannot use kurtosis as a common estimator for non-Gaussianity. Alternatively, we estimate the not uniquely defined peakedness for evaluating the "difference" from the Gaussian distribution. According to the literature, several such measures have been defined based on different definitions or interpretations of the term "peakedness" (see Brys et al. 2006; Schmid and Trede 2003). Among them we have selected 3 possible robust estimates for peakedness which are all based on quantiles described by Schmid and Trede 2003. Thus, they cannot only be used for samples but also for symmetric distributions with known parameters. These contain the T measure,

$$T = \frac{quant(0.875) - quant(0.125)}{quant(0.75) - quant(0.25)} \tag{7.8}$$

the P measure

$$P = \frac{quant(0.975) - quant(0.025)}{quant(0.875) - quant(0.125)} \tag{7.9}$$

and the L measure which is the ratio of the previous two measures

$$L = P \cdot T = \frac{quant(0.975) - quant(0.025)}{quant(0.75) - quant(0.25)}. \tag{7.10}$$

Figure 7.3 plots the P and T measure within the relevant interval for the degrees of freedom ν. Both functions show a similar behaviour and a larger slope for smaller values of ν, as can be expected. For small degrees of freedom the student's t distribution becomes very heavy-tailed and the distance of the quantiles from the Gaussian quantiles grows exponentially. Thus, these measures allow us to discern very well between the "interesting" distributions with degrees of freedom less or equal to 4 and t distributions with larger degrees of freedom and the Gaussian distribution. The additional advantage of the T measure would be its higher breaking point in the view of classical robustness which however is only relevant for sampe based estimation of 'peakedness'. The P and L measure on the contrary are particularly useful for our estimation based on the known distributions' exact quantiles. For discerning between heavy-tailed distributions it is important to consider the difference between more 'extreme' persentiles, such as 0.975 and 0.025 as opposed to 0.875 and 0.125. As we calculate our measure based on the estimated degrees of freedom and the exact quantiles of the distributions, this is of no effect for us.

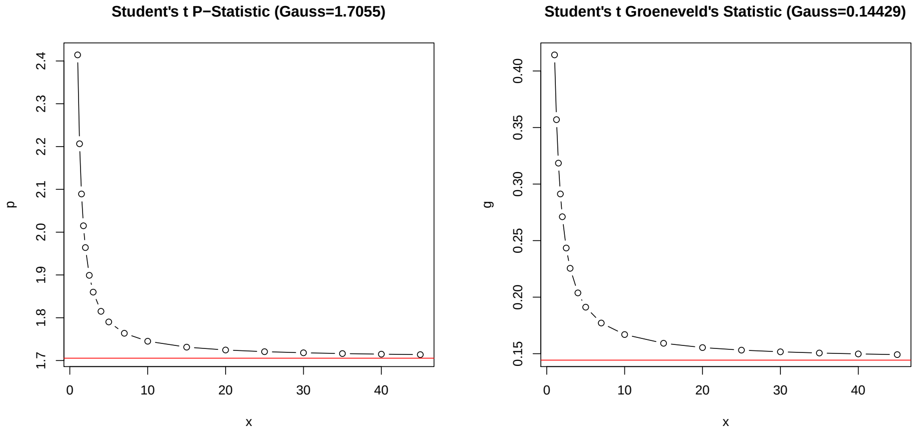


Figure 7.3: The "peakedness" estimator dependent on the degrees of freedom. The red line marks the values for the Gaussian. To the left is the P measure, to the right the T measure.

Applying loss functions for dealing with the label switching problem has first been suggested by Celeux et al. 2000, while Frühwirth-Schnatter 2001 has suggested to introduce identifiability constraints. We have adapted the ideas and not introduced a loss for formal Bayesian posterior inference, but rather a transformation of the variables ν_j. This approach goes along the line of the approach which Frühwirth-Schnatter 2011 suggested in their more recent work. Without the introduction of the "peakedness" measure for non-Gaussianity we could not perform posterior inference of the degrees of freedom parameters ν_j. As we have chosen to include the Gaussian model ($\nu = \infty$) into our model instead of a student's t approximation ($\nu = \nu_{max} < \infty$), averaging over the ν_js is impossible. As soon as a chain has jumped to the Gaussian model once, the result would be infinite. Untangling the chains intermixed by label switching would pose an additional challenge. This we would need to tackle applying established algorithms which however would introduce additional errors.

As our goal is to identify the differential expression behaviour in the current data set, not to perform predictions based on it, we need not identify the mixture components and their respective weights. Instead, we only have to deal with identifying the "components" or groups of genes with similar peakedness behaviour, which is a much simpler one-dimensional problem. By calculating the gene-wise average peakedness, we gain a reasonable measure for non-Gaussianity based on the P measure

$$P - peakedness(g) = \sum_{i=1}^{N} \frac{quant(0.975; \nu_i^{(g)}) - quant(0.025; \nu_i^{(g)})}{quant(0.875; \nu_i^{(g)}) - quant(0.125; \nu_i^{(g)})} \quad (7.11)$$

for $g = 1, \ldots, G$ where $\nu_i^{(g)}$ is the degrees of freedom parameter assigned to gene g in the i^{th} sample run. For this measure we can for example adopt the approach of defining identifiability constraints a posteriori to separate the noise groups. Our measure is also consistent with Monte Carlo theory and thus provides an asymptotically unbiased estimator.

Our peakedness-based measure also allows us to represent our trust or lack thereof in hypothesis testing results for very heavy-tailed components, which can contain many outliers. Using a properly chosen non-Gaussianity measure, we are able to deal with cases where the over-dispersion most likely originates from errors during the measurement process. Especially for experiments with very few arrays available, this might be a reasonable approach for selecting genes among the top-ranked results. Based on this scheme it is possible to down-weight the respective observations and thereby express lack of trust. For experiments with large sample sizes the genes following t distributions however do not result from single erroneous arrays, but might rather indicate systematic over-dispersed behaviour which can originate from biological processes or seriously error-prone laboratory work. Both scenarios are of interest for the data analyst.

7.2.11 A new approach to microarray quality control

For the purpose of microarray quality control, we propose a novel empirical Bayes ansatz to test the influence of single arrays on noise behaviour. One reason for outliers in microarray analysis is that several genes on a single array are affected by an array-specific problem during conducting laboratory work. Examples would be blotches of dye or scratches on the glass. Commonly, these problems are only found when performing several tedious steps of quality control and bioinformatic preprocessing, as they will not be detected during the analysis, if overlooked in the preprocessing. Our intention is to find the array(s) most likely responsible for extreme noise behaviour, if such arrays occur in the experiment. The noise structure allows us to isolate the most affected genes, tracing them back to the responsible array(s).

In order to compare the most probable posterior degrees of freedom against the Gaussian for each array and gene, we suggest the Bayes Factor

$$BF_{n,g} = \frac{\mathbb{P}[y_{n,g}|\nu = \hat{\nu}_g, \beta_g = \hat{\beta}_g, \lambda = \hat{\lambda}_g]\mathbb{P}[\nu = \hat{\nu}_g]}{\mathbb{P}[y_{n,g}|\nu = \infty, \beta_g = \hat{\beta}_g, \lambda = \hat{\lambda}_g]\mathbb{P}[\nu = \infty]}. \tag{7.12}$$

This ratio can be interpreted in the following way: A Bayes Factor $BF_{n,g}$ close to one implies an equal probability of Gaussian and t noise for this array's measurement. Values close to 1 favour the Gaussian, values far greater than 1 the t alternative.

Our aim is to find a value δ, to the effect that all (n, g) with $BF_{n,g} > 1+\delta$ drive the noise towards t. As this will be the case for the majority of observations on microarrays, we intend to look only at the most extreme cases. For reasons of practicality, we choose the most extreme 5 to 10%, or a percentage depending on the weight of the most extreme component (cf. Section 7.4.1). Afterwards, we count how many of these observations correspond to which array. For a properly conducted experiment it is expected that these counts are approximately uniformly distributed among the arrays.

As the distribution of Bayes Factors is extremely data dependent, this approach requires empirical Bayes methods rather than a fully Bayesian model, in which the prior is completely independent of the data. In correspondence to Pearson's χ^2 statistics, we have constructed the following test: Our basic assumption is that the observed counts of extreme values of the N arrays are multinomially distributed with identical probability $1/N$ under the null hypothesis, $(N_1, \ldots, N_N) \sim \mathcal{M}_{N_{sel},\pi}$. In a conjugate distribution scenario we assume that the probability π follows a Dirichlet distribution, $\pi \sim Dir(\alpha_0, \ldots, \alpha_0)$. The choice of the prior parameter α_0 of this Dirichlet distribution is highly influential. Thus, we employ the outcomes of a corresponding χ^2 test as an orientation for this value, which has to depend on the number of arrays N. We decide for a value of $\alpha = 2.5 \cdot N$, as it has shown the best behaviour in a series of tests. In order to test the hypothesis of uniformly distributed extreme values among the arrays, we apply the Savage Dickey Density Ratio (Dickey and Lientz 1970). The Savage-Dickey density ratio is a special type of Bayes factor required if one of the considered hypothesis in he Bayesian testing scheme is a point hypothesis of a continuous distribution.

$$\begin{aligned} H_0: \quad \pi &= (1/N, \ldots, 1/N) \\ H_A: \quad \pi &\neq (1/N, \ldots, 1/N) \end{aligned} \tag{7.13}$$

$$SDR = \frac{p(\pi = (1/N, \ldots, 1/N) | (\alpha_1^*, \ldots, \alpha_N^*))}{p(\pi = (1/N, \ldots, 1/N) | (\alpha_0, \ldots, \alpha_0))} \tag{7.14}$$

A Savage density ratio above 1 represents evidence in favour of the null hypothesis, whereas values less then 1 show evidence against H_0. Our first hypothesis, termed H_0 in analogy to the χ^2 test which we compare it to, represents equal amounts of extreme genes on all arrays. This is the optimal case of no array drawing the analysis. The alternative hypothesis we are interested in is the case when at least one array shows a different relative amount of extreme genes compared to the others, thus influencing the analysis and not fulfilling the minmum quality. Detecting such arrays allows to reconsider ones results obtained with the model in which they are included to approaches where observations of this particular array are downweighted or removed completely from the analysis which often is not possible or reasonable due to the small sample sizes, compare for example 6 arrays in total for the spike-in data. This way, our approach for quality control allows for an improved analysis of the most extreme microarray data.

7.3 Results for artificial data

Applying our method to artificial data, we find that our algorithm is specifically able to identify extreme noise situations, i.e. Gauss mixed with very heavy-tailed student's t components. In case of high degrees of freedom t distributions our method tends towards favouring the simpler Gaussian model, penalising the computationally and memory-intense estimation of the unnec-

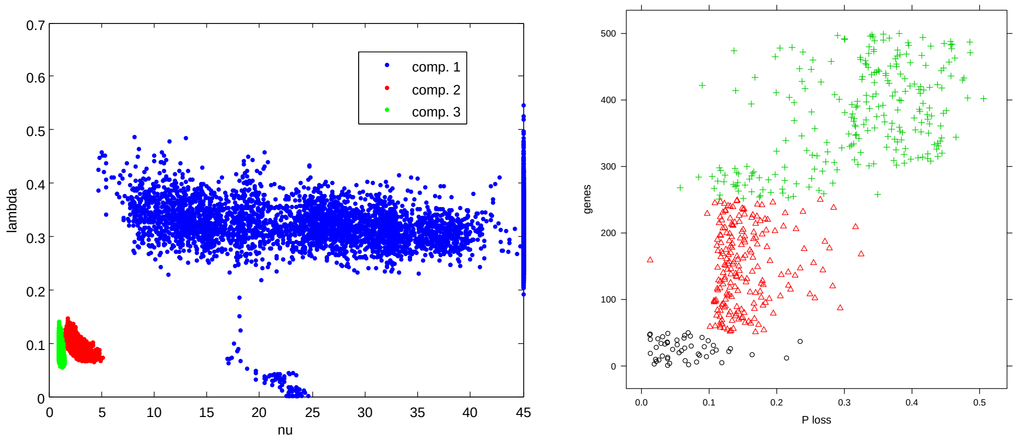


Figure 7.4: The projection of the posterior distribution on the ν-λ plain (left) and the P measure plotted against the posterior probability of differential expression (right) for 10% Gaussian data, 50% t_{10} data and 40% t_1 data. Although no obvious intermediate t_{10} component forms in the ν-λ space, the ”p loss” measure captures the known noise behaviour of the 3 components, visualised by the different plotting symbols and colours in the right graph. Appendix A of the supplement includes further scenarios.

essary rescaling parameters. This property makes it favourable to use for memory-intense purposes, e.g. in bioinformatics, compared to mixture models which contain only student’s t distributions, such as Stephens’ method (cf. 6). For the test data sets described above, we observe that the algorithm is always able to discern between clearly distinct noise components, such as the Gaussian from t_4 and t_1 or t_1 from t_4, t_7 and t_{10}. These results are summarised in Table 7.1.

Separating the t_{10} from the t_4 distribution or the Gaussian depends on the setting of the data and the mixture weights, as the algorithm favours smaller and simpler models. For high degrees of freedom t distributions our method favours the simpler Gaussian distribution, thus, splitting t_{10} observations between the heavy-tailed t_4 or t_1 and the normal component. When mixing t_1 and t_{10} data with Gaussian data, a simpler model with 2 components would apparently suffice, when looking at the (ν-λ) graph. In such cases, the method splits the t_{10} observations between the heavy-tailed and the normal component in such a way that the last component is only fluctuating regarding its weight and degrees of freedom. This behaviour is reassuring, as we learn that unnecessary complexity in form of the memory- and computationally intense estimation of unnecessary rescaling parameters is automatically penalised by the model. This property is valuable for analysing large amounts of data, such as microarray experiments for which the method is ultimately intended.

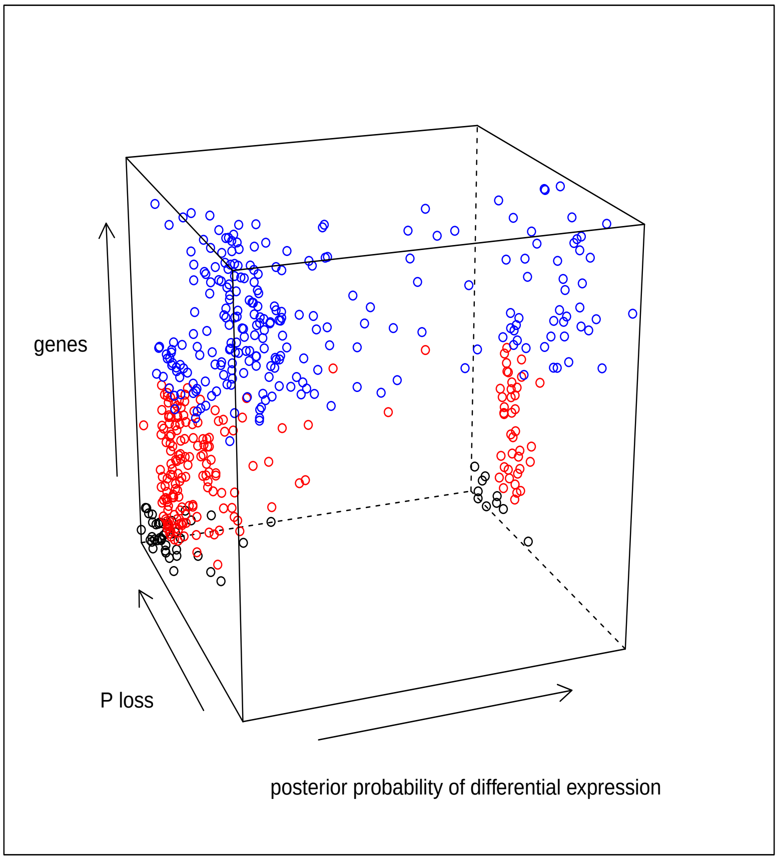
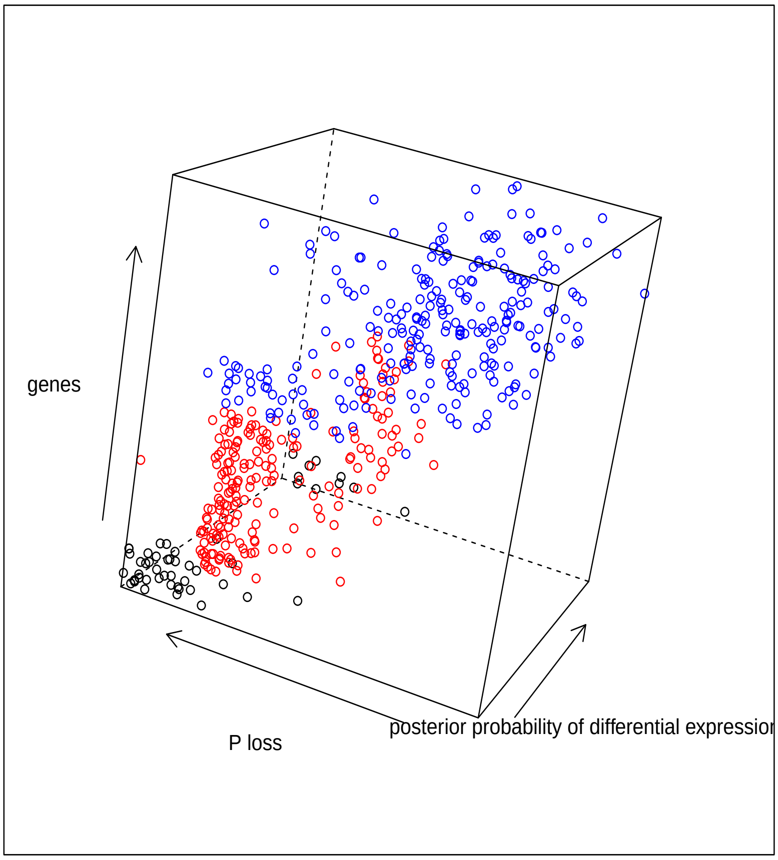


Figure 7.5: Two three-dimensional visualistions of the artificial data set with 10% Gaussian data, 50% t_{10} data and 40% t_1 data which include the genes, the posterior P loss function and the posterior probability of differential expression.

Figure 7.4 visualises our defined peakedness measure applied to the artificial gene data sets. We show how the noise components drawn with the two noise coordinates (ν, λ) split up in the plain of noise coordinates (ν, λ). Clearly at least one heavy-tailed student's t component forms. In the setting of two heavy-tailed student's t components, t_4 and t_1, both components can be seen clearly in the two-dimensional projection of the noise



space. The visualisation of the non-Gaussianity measure also identifies these components as can be seen in the right graphic of figure 7.4. When we mix t_4 and t_{10} data with Gaussian data, the t_{10} data are split up between the t_4 and Gauss components, which explain the majority of the data, while the 3^{rd} component only induces label switching, but has hardly any weight. Here, the simpler model with 2 components would apparently suffice. However, the non-Gaussianity measure discriminates 3 components, corresponding to our original setting of Gauss, t_1 and t_{10} data. This demonstrates the advantage of our peakedness measure for the identification of noise model components. Thus, the advantage of our peakedness measure becomes clear for the identification of noise model components, as it also considers the individual weights of the components for each 'virtual' gene.

We have also used the test data sets with 5 different noise settings, cf. table 7.1, for estimating the general sensitivity and specificity of the algorithm. Both are measures of performance of the algorithm, which are based on the number of true and false positives and negatives, respectively. Given the performance of the MCMC runs, we can conclude that for the identification of differential expression behaviour (the biological analysis goal) the sensitivity lies between 0.90 and 0.97 for the cumulative analysis of 5 parallel chains, while the specificity amounts to 0.97 to 0.99. For identifying the correct noise model (Gaussian vs. non-Gaussian) the specificity of the algorithm is 0.98, while the sensitivity is 0.9. However, we could observe that some chains perform less well than the others and draw this assessment. Thus, a careful analysis of convergence is essential before including MCMC output in the final summary. By the means of the coda package we could identify chains with worse convergence behaviour. Excluding them from the calculations substantially improved the results.

7.4 Bioinformatical analysis

7.4.1 Microarray data

In addition to artificial data, we analyse different microarray data sets. The non-Gaussianity measure based on peakedness helps us to identify non-Gaussian noise. This behaviour may originate from a biological sub-process or from the execution of the experiment. In order to identify such laboratory effects we analysed the ”Golden Spike” experiment. Choe et al. 2005 performed a Spike-In experiment, where all genes’ behaviour w. r. t. differential expression is known. The only noise in the experiment originates from laboratory procedures. The genetic material used is taken from flies.

Unlike the Spike-in data there is no ”gold standard” data set available for any biological microarray experiment. We chose 4 large data sets from the Gene Expression Omnibus (Edgar et al. 2002b) data base for the analysis. The data set with GEO ID GDS2960 analyses marfan syndrome in humans with 101 microarrays. The original study has been conducted by Yao et al. 2007b. The second data set has the GEO ID GDS1375 and includes 70 arrays in the experiment. Talantov et al. 2005b studied types of human melanoma and their genetic differences. In this study the authors identified certain marker genes for differing between malignant and non-malignant melanoma types. Our third data set, GEO ID GDS531 deals with myeloma and contains 173 arrays. Tian et al. (Tian et al. 2003) performed the original study on cells from human bone marrow. The fourth study, GEO ID 2946, contains 15 arrays and thus far less than the first three ones. However this number is more representative of the typical number of sample sizes available for microarray experiments. Li et al. 2008b studied obesity in rats subjected to different diets.

Before the gene expression analysis, we performed preprocessing as it is typically conducted for such microarray data, cf. Speed 2003 and explanations in chapter 3. GDS2960 is performed on invitrogen and uses RPG3.0 for preprocessing the arrays. All other experiments use Affymetrix platforms, the data were read as MAS5.0 and then normalised. Normalisation in this case refers to the bioinformatical normalisation which aims for removing background effects of the measurement in order to make the arrays' measurements better comparable, cf. the discussion of the difference between the statistical and bioinfomratical view of normalisation. In statistics, the idea behind normalization is transforming data towards the normal distribution. In bioinformatics however it stands for transforming the data to remove unwanted effects introduced during the laboratory process and making it easier to handle. As we wish to work in a framework close to normal distribution settings, we want normalisation to work in the original sense of the word: transform towards normality.

Figure 7.6 shows how the variance stabilising normalisation (vsn) by Huber et al. 2002 transforms towards normality, while other standard normalisations, loess or quantile, keep the underlying highly skewed structure of the unnormalised data. Therefore, it is reasonable to utilise vsn normalised data for our further analyses and observe loess for comparison in parallel. However, we can show with the spike-in data that the skewness behaviour enforces a trend towards heavy-tailed distributions, when only few (2-4) components are fitted, which is not observed for vsn data or more components.

Results for Microarray Data

First, we analyse the spike-in data with known differential expression behaviour, where the noise only originates from laboratory work. Here, we fit

Table 7.2: The total weights of Gaussian mixture components, which is the sum of component weights, if more than one component is Gaussian, for different data sets (marked by their GEO ID number), normalisations (vsn and quantile) for models with various numbers of components. The technical spike-in data behaves differently than the biological data, with only 0-25% properly modelled by Gaussian distributions.

GEO ID	normalisation	mixture components			
		2	3	4	5
GDS2960	vsn	0%	0%	15%	13%
	quantile	0%	0%	7%	10%
GDS1375	vsn	0%	7%	8%	8%
	quantile	0%	0%	0%	0%
GDS531	vsn	0%	0%	28%	45%
	quantile	0%	0%	0%	0%
GDS2946	vsn	0%	8%	24%	25%
	quantile	0%	0%	0%	0%
golden spike	vsn	100%	100%	100%	100%
	quantile	0%	0%	33%	89%

mixture models with 2 and 3 noise components, being the most plausible models indicated by the algorithm with variable number of components, and 4 and 5 component models as well. In particular, we look for differences between spike-in and biological data, which has not been found in a previous systematic study of noise behaviour, where Posekany et al. 2011 have shown that the overall noise of the ”golden spike” data is as heavy-tailed as for real microarray data sets. For the vsn-normalised spike-in data, the mixture model identifies only Gaussian components, in which one component with very low precision fits the data containing the most noise. This is plausible, as Student’s t distributions can be approximated by mixtures of Gaussians, indicating a regular noise behaviour of laboratory factors. However, this does not imply that over-dispersion is generally not caused by problems during laboratory work, but it shows that, for well-conducted experiments, the technical noise behaves regularly and that the influence is typically Gaussian. When using the highly skewed quantile-normalised data, we can still observe a trend of models with 4 or more components favouring normal components, which appears for none of the biological data sets. Thus, there is no trend of a differing noise behaviour between the genes, the normal components apparently model the single sutdent's t noise overall fitting for this data.

Applying the variable component algorithm to the biological data sets, small models with 2 or 3 components should suffice, but as with spike-in data we fit models with 4 and 5 components in addition. Normalising with vsn has the positive effect of reducing over-dispersion. However, heavy-tailed components still remain in the data, but Cauchy-distributions do not dominate the model; a behaviour we observe for quantile normalisation. Contrary to this, the skewness of quantile normalised data enforces a trend towards heavy-tailed distributions, when fitting only few (2-4) components, which is not observed for vsn data or more components. This observation is consistent with previous findings (Posekany et al. 2011), where vsn normalised data follow overall t_4 distributions, whereas quantile normalised data require a Cauchy (t_1) distribution. Thus, choosing a preprocessing methodology for microarray data, which includes normalisation, is highly influential on the analysis and has to be taken into account. Apparently, a more differentiated noise behaviour with far more extreme observations comes to the surface when the genes' concentrationis not prepared in the laboratory but stems from underlying biological processes the microarray experiment wishes to untangle and observe.

Our main finding is that at least one heavy-tailed Student's t component explains the majority of the noise, independent of the fitted number of components or the normalisation. This marks a striking difference compared to the spike-in data as well as to Novak et al. 2006b's findings that 85 to 95% of the microarray data follow a normal distribution, although they have remarked that the extreme data in microarray experiments have quantiles similar to Student's t distributions. Our assessments of the estimated weights of components reveals that the relation seems to be the other way around: Only 5-25 % of the data follow a normal distribution, with numbers depending on data set and normalisation (cf. table 7.2). In models with more than 3 components, Gaussian components form small subcomponents. However, this observation might occur due to the Gaussian component(s) mixing together to form a single more heavy-tailed student's t component. This would be mean that the Gaussian components occurring for large mixture sizes originate from overfitting rather than an interpretable process, as an inherent penalty against fiting student's t components exits when Gaussian components would suffice. For two and three component models, which have been found to be most reasonable and interpretable, only Student's t components are fitted. This confers to previous studies by Hardin and Wilson 2009 and Posekany et al. 2011, who have observed that Gaussian distributions are overall unfitting for microarray data. In particular, this finding marks a difference between the laboratory-based spike-in data and biological data which apparently introduces an inherently more heavy-tailed behaviour from the molecule-generating processes in the cell which is not present in laboratory data.

When applying the peakedness score, we can observe in figure 7.7 for GDS2960 and GDS531 that the most heavy-tailed component, including about 10-15% of the data, identifies noisy genes with hardly any differential expression behaviour. This clear split only occurs when fitting 4 or more components, as for less components too many genes are pooled in a single component to allow more detailed identification. Therefore, our measure of non-Gaussianity allows us to find the disturbing, over-dispersed genes and thus provides a useful tool for separating them from the rest, which is valuable for considerations regarding the reliability of genes in gene expression analysis.

7.4.2 Array Quality control

As the majority of microarray genes does not show Gaussian noise behaviour, representing the severe over-dispersion, which data analysts have to struggle with, we do not look for all genes and arrays introducing a trend towards t distributions, but only count the most extreme ones. Due to the complexity of microarray data and its unknown true behaviour, we will take the generated data as the gold standard for testing our method of microarray quality control.

Based on our test data sets, we are able to show that our ansatz recognises scenarios, in which the noise is split equally among the arrays as well as scenarios with single arrays containing the majority of the extremely noisy genes. For our test data, a compare ison of the outcomes of a regular χ^2 test

Table 7.3: We compare the outcomes of a regular χ^2 test and the Bayesian test, using Savage Density Ratio (SDR) for the microarray-like test data consisting of 10% normally distributed data and 90% Student's t data, which is split up according to the first column. For equally split extreme noise, we observe no significant findings, provided by the SDR test (or the χ^2 test). In the alternative scenario we randomly select 50% or 75% of the genes containing the most extreme 10% of Bayes Factors. The extreme values of these genes are accumulated on a single array, a scenario which does not affect the noise estimation or differential expression assessment at all. Here, our test detects this unbalanced situation reliably.

	equal extremes		50% most extreme		75% most extreme	
40%- 50%	χ^2 test	SDR	χ^2 test	SDR	χ^2 test	SDR
$t_4 - t_{10}$; 5%	0.15	2.64	0.028	0.42	7.2e-11	7.4e-07
$t_4 - t_{10}$; 10%	0.13	2.07	0.0024	0.058	4.0e-33	1.6e-19
$t_4 - t_{10}$; 15%	0.47	12.56	0.015	0.22	3.4e-19	2.1e-12
$t_4 - t_1$; 5%	0.69	29.26	0.0020	0.043	1.5e-14	3.8e-09
$t_4 - t_1$; 10%	0.22	4.03	0.000056	0.0034	2.8e-29	9.5e-18
$t_4 - t_1$; 15%	0.40	9.19	0.0013	0.030	1.6e-13	3.3e-09
$t_7 - t_1$; 5%	0.33	8.27	0.056	0.82	1.2e-14	4.3e-09
$t_7 - t_1$; 10%	0.72	33.44	0.00055	0.018	1.3e-38	2.0e-22
$t_7 - t_1$; 15%	0.88	58.45	0.014	0.22	1.5e-20	2.9e-13
$t_{10} - t_1$; 5%	0.53	15.83	0.00029	0.010	1.1e-03	5.5e-02
$t_{10} - t_1$; 10%	0.07	0.90	0.0046	0.078	1.7e-04	7.4e-03
$t_{10} - t_1$; 15%	0.18	2.85	0.0048	0.10	4.3e-03	1.1e-01

and the Bayesian test is conducted. The Bayesian testing approach uses the Savage Density Ratio (SDR), as described in section 7.2.11, and applies it to the microarray-like test data consisting of 10% normally distributed data and 90% Student's t data. For extreme noise data which is by generation equally distributed among all arrays, we observe no significant findings, provided by the the χ^2 test and SDR test. This scenario has been used for sensitivity analysis of choice of α_0 which led to the conclusion that a data-independent fully Bayesian prior with any choice of hyperparameters is inferior to an empirical Bayesian choice of data-dependent prior. As alternative scenarios we randomly select 50% or 75% of the genes containing the most extreme 10% of gene and array-dependent Bayes Factors $BF_{n,g}$, cf. formula (7.12). These extreme values simulated for the artificial genes are accumulated on a single array. This scenario does not affect the noise estimation or differential expression assessment at all, but it should be detected by a measure for determining microarray quality control. In the artificial scenarios, our test detects this unbalanced situation reliably as the Bayes factors show that the hypothesis of unequal weights is several times as probable as the balanced hypothesis. Table 7.3 includes the results of our empirical Bayes test and shows the analogy with the classical χ^2 test. In all shown cases with equally split noise, the test recognises that the observed behaviour is not due to a single noisy array, but stems from the underlying behaviour of the genes or sources which are common to all arrays. If a certain amount of extremely noisy genes has influential values, located on a single array, our approach identifies such an unbalanced situation correctly. The sensitivity to 'unbalancedness' is adjusted by the prior of α_0. The data-dependent choice of prior for α_0 partially allows for a semi-automatic approach, while careful adjustment of priors can always lead to better results than standard choices as for any Bayesian computational inference scheme. For application tuning the parameters of he test and reweighting the Savage density ratio with an informative loss scheme unlike the $0-1$ loss which we currently apply. This can lead to an increase or decrease of the test's sensitivity to the percentage of 'misbehaving' genes on a single array before an unbalanced situation is found to be significant. Our findings have the potential of leading to improvement bioinformatical analysis of microarray data by adding a quality control option to the differential expression assessment.

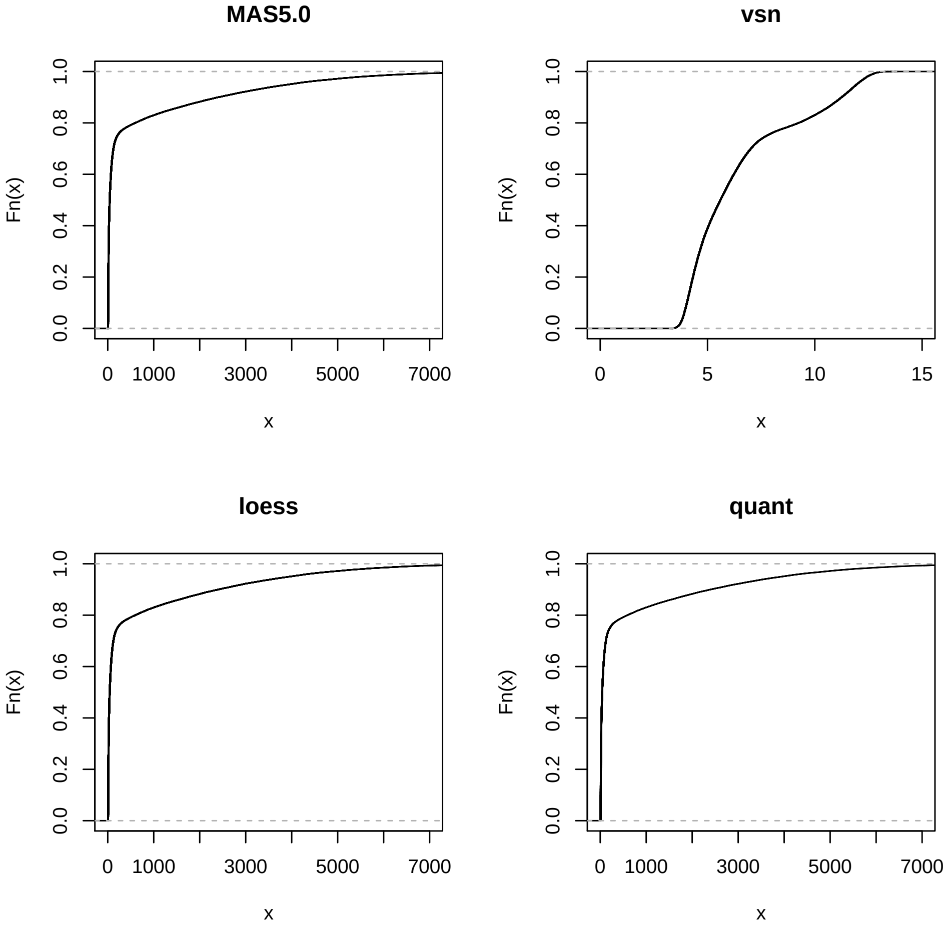


Figure 7.6: Empirical cumulative distribution functions of the data accumulated over all genes and arrays for different normalisations of the "golden spike" data. In the top left corner the unnormalised data, in the top right the vsn normalised and in the bottom, loess and quantile normalised data.

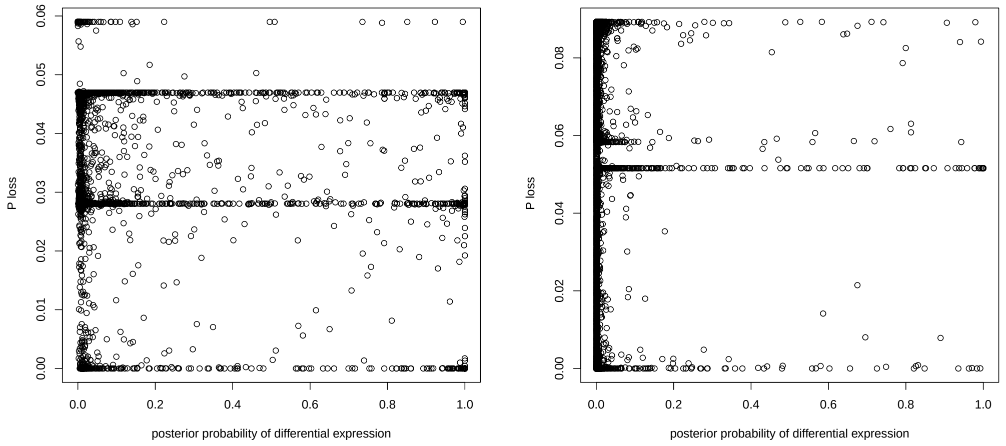


Figure 7.7: Graphs of the posterior probability of differential expression, plotted against the peakedness measure for the GDS2960 and GDS531 data for 4 component models. These scatter plots relate the biological variable of interest for evaluating the importance of genes to our measure of non-Gaussianity, denoted as "P loss". Genes on the right side have low probability of differential expression and no biological relevance, whereas genes on the left side are the "targets" of the experiment. The most heavy-tailed component contains genes with low probabilities of differential expression. A Gaussian component with loss close to 0 contains some biologically relevant genes, while the majority of differentially expressed genes belongs to the Student's t component with low degrees of freedom.

Chapter 8

Original ideas of this work

To summarise the work presented in this thesis, we will now point out all the novel approaches, implementations and findings in this last chapter. To make conception easier, all the important points will be divided up into three main categories, namely statistical modelling approaches, computational implementation and bioinformatical findings.

- *Statistical approaches*
  The two hierarchical Bayesian models in chapters 5 and 7 both present hand-tailored methods specifically designed to deal with microarray data. The basic idea of robust Bayesian likelihood estimation is related to Bayesian model selection, i. e. the a posteriori best fitting model is selected. Model 5.1 applies this idea for the comparison of noise models by considering likelihood functions. As student's t distributions are able to keep the symmetry of residual models and a proximity to the normal distribution, they provide an excellent choice for dealing with overdispersion.

  Chapters 5 and 7 provide different approaches for prior distribution settings of the degrees of freedom parameter. The flexible, yet discrete set is a unique approach which was first developed by Posekany 2009. The continuous uniform distribution approach has been extensively used and studied, compare for example the implementation by Frühwirth-Schnatter 2006. However, in this setting the novel scheme of simultaneously considering normal and t models leads to the advantage of avoiding the known extreme sensitivity of t distributions to the cut-off of the prior over the degrees of freedom parameter. Alternative schemes do not consider the Gaussian distribution as an option to 'jump to', but approximate it by student's t distributions with high degrees of freedom. However, such alternatives do not allow a clear distinguishing between the distributions, whereas our scheme, provided in chapter 7, combined with our peakedness measure clearly separates the different distributions w. r. t. their respective tail weight. The alternative approach of identifying tail weights based on similar schemes, e. g. for mixtures of normals, would not lead to straight-forward results, as on the one hand the curtosis to be approximated does not exist, while on the other hand the peakedness is the same for all normal distributions. Our method of applying measures of 'peakedness' is a unique approach, as it discriminates student's t distributions with degrees of freedom too small to reasonably determine an estimate for curtosis. Our approach's advantage lies in providing a flexible measure for discerning the symmetric distributions with different tail behaviour considered in this work.

- *Computational implementations and approaches*

  Our algorithmic contribution lies in the implementation of the novel hybrid MCMC sampler for model 7.1, which includes a reversible jump step between student's t and normal distributions in the microarray analysis ANOVA setting, in particular in the mixture analysis setting. First, we performed a detailed analysis of the algorithm's 'sanity', i. e. sampling in known cases from a proper posterior and being able to provide useable results within a reasonable amount of time. Second, we applied the algorithm to a selected sample of microarray data sets.

  As is generally known, mixture algorithms are prone to specific problems, such as label switching, due to the inherent lack of identifiability of the respective components, cf. chapter 6. In addition to identifying distributions, the above-mentioned peakedness measure allows us to overcome the label switching problem by performing posterior inference over the not label-related 'peakedness' instead of the label-dependent variables, the label $S_{n,g}$ the precisions λ_j, degrees of freedom ν_j and rescaling factors $\varphi_{n,g,j}$.

- *Bioinformatical findings*
  Previously, such a large-scale systematic study of microarrays' noise behaviour as Posekany et al. 2011 has never been conducted in the realm of bioinformatics. The aim of our study was to compare different approaches of modelling noise in linear regression settings by performing Bayesian model selection based on criteria of robustness. Our findings largely supported the heavy-tailed distributions' model to provide a better fit for the noise in microarray data than the normal distribution. Our unpreceded study of the quantitative effects on secondary bioinformatical analyses, such as gene ontologies, led us to the conclusion that the errors made by wrong model assumptions can lead to large errors in biological conclusions and findings. In other words, the whole process of bioinformatical analysis is very sensitive to unfitting choices of model components, in particular of the likelihood function.

  In order to test existing assumptions about the relative amount of data following the normal distribution, the mixture model (7.1) was introduced. The mixture of normal and t distributions has always been meant to not only provide a better distribution approximation, but also to allow for an interpretation of its components. This second feature would be lost, if we fitted the model with mixtures of normal distributions with enough components. A possible way to reasonably use the information about the 'peakedness' of our model's associated components lies in array quality control, in which a single array with genes not following the behaviour provided by the other arrays might be found. Alternatively, genes differing in their more or less noisy behaviour could be singled out from the rest. For future investigations it would be an interesting possibility to explore in detail, whether the reason for the observed effects stems from problems resulting from laboratory work or from underlying biological processes.

All in all, out thesis had the foremost aim to unite Bayesian statistical modelling with the field of bioinformatics, an application of great importance for modern scientific research. Challenges stemming from the irregularly behaved data as well as the complicated computational implementation of the two different models were overcome by relying on specific ressources such as the Vienna Science Cluster. We are sure that we have been able to provide profound results on which further investigations can be built.

Bibliography

[1] M. Affara et al. "Understanding Endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal?" In: *Philosophical Transactions of the Royal Society B* 362 (2007), pp. 1469–1487.

[2] F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. "FatiGO: a web tool for finding significant association of Gene Ontology terms with groups of genes". In: *Bioinformatics* 20 (2004).

[3] M. Ashburner et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." In: *Nature Genetics* 25 (2000), pp. 25–29.

[4] K. Bae and B. Mallick. "Gene selection using a two-level hierarchical Bayesian model". In: *Bioinformatics* 20 (2004), pp. 3423–3430.

[5] P. Baldi and A. Long. "A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes". In: *Bioinformatics* (2001).

[6] J. Banfield and A. Raftery. "Model-Based Gaussian and Non-Gaussian Clustering". In: *Biometrics* 49.3 (1993), pp. 803–821.

[7] J. Berger. *A Catalog of Noninformative Priors*. Department of Statistics, Purdue University.

[8] J. Berger et al., eds. *Bayesian Robustness*. IMS Lecture Notes, 1995.

[9] James O. Berger. "An overview of robust Bayesian analysis". In: *Test* 3 (1994), pp. 5–124.

[10] J. Bernardo and A. Smith. *Bayesian Theory*. Series in Probability and Statistics. Wiley, 2000.

[11] J. Besag et al. “Bayesian Computation and Stochastic Systems”. In: *Statistical Science* 10.1 (1995), pp. 3–41.

[12] C. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[13] E. Blalock et al. “Incipient Alzheimer’s disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses”. In: *Proceedings of the National Academy of Sciences* 101.7 (2004), pp. 2173–8.

[14] B. Bolstad et al. “A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance”. In: *Bioinformatics* 19.2 (2003), pp. 185–193.

[15] G. Box. *Robustness in the Strategy of Scientific Model Building.* Tech. rep. University of Wisconsin–Madison, 1979.

[16] G. Brys, M. Hubert, and A. Struyf. “Robust measures of tail weight”. In: *Computational Statistics & Data Analysis* (2006), pp. 733–759.

[17] D. Cameron et al. “Gene expression profiles of intact and regenerating zebrafish retina”. In: *Molecular Vision* 11 (2005), pp. 775–91.

[18] O. Cappe, C. Robert, and T. Ryden. “Reversible Jump, Birth-and-Death and More General Continuous Time Markov Chain Monte Carlo Samplers”. In: *Journal of the Royal Statistical Society. Series B* 65.3 (2003), pp. 679–700.

[19] G. Celeux, M. Hurn, and C. Robert. “Computational and Inferential Difficulties with Mixture Posterior Distributions”. In: *Journal of the American Statistical Association* 95.451 (2000), pp. 957–970.

[20] S. Choe et al. “Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset”. In: *Genome Biology* 6.R 16 (2005).

[21] M. Cowles and B. Carlin. “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review”. In: *Journal of the American Statistical Association* 91.434 (1996), pp. 883–904.

[22] S. Crawford. “An Application of the Laplace Method to Finite Mixture Distributions”. In: *Journal of the American Statistical Association* 89.425 (1994), pp. 259–267.

[23] J. De Haan et al. “Robust ANOVA for microarray data”. In: *Chemometrics and intelligent laboratory systems* 98 (2009), pp. 38–44.

[24] G. Dennis et al. “DAVID: Database for Annotation, Visualization, and Integrated Discovery”. In: *Genome Biology* 4.3 (2003).

[25] J. Dickey and B. Lientz. “The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain”. In: *Ann. Math. Stat.* 40.1 (1970), pp. 214–226.

[26] J. Dinneny et al. “Cell identity mediates the response of Arabidopsis roots to abiotic stress”. In: *Science* 320.5878 (2008), pp. 942–5.

[27] K. Do, P. Müller, and F. Tang. “A Bayesian mixture model for differential gene expression”. In: *Applied Statistics* 54.3 (2005), pp. 627–644.

[28] D. van Dyk and T. Park. “Partially Collapsed Gibbs Samplers: Theory and Methods”. In: *Journal of the American Statistical Association* 103.482 (2008).

[29] R. Edgar, M. Domrachev, and A. Lash. “Gene expression omnibus: NCBI gene expression and hybridization array data repository”. In: *Nucleic Acid Research* 30 (2002), pp. 207–210.

[30] R. Edgar, M. Domrachev, and A. Lash. “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.” In: *Nucleic Acid Research* 30(1) (2002), pp. 207–210.

[31] D. Fink. *A Compendium of Conjugate Priors*. Tech. rep. Montana State Univeristy, 1997.

[32] S. Frühwirth-Schnatter. “Dealing with label switching under model uncertainty”. In: *Mixture estimation and applications*. Ed. by K. Mengersen, C. Robert, and D. Titterington. Wiley, 2011, pp. 193–218.

[33] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer-Verlag, 2006.

[34] S. Frühwirth-Schnatter. “Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models”. In: *Journal of the American Statistical Association* 96.453 (2001), pp. 194–209.

[35] S. Frühwirth-Schnatter and S. Pyne. “Bayesian Inference for finite mixtures of univariate skew-normal and skew-t distributions”. In: *Biostatistics* 11.2 (2011), pp. 317–336.

[36] X. Gao and P. Song. “Nonparametric tests for differential gene expression and interaction effects in multi-factorial microarray experiments”. In: *BMC Bioinformatics* 6 (2005), p. 186.

[37] A. Gelman. “Prior distributions fror variance parameters in hierarchical models”. In: *Bayesian Analysis* 1.3 (2006), pp. 515–533.

[38] A. Gelman et al. *Bayesian Data Analysis*. Chapman & Hall, 2003.

[39] Robert C Gentleman et al. “Bioconductor: Open software development for computational biology and bioinformatics”. In: *Genome Biology* 5 (2004), R80. URL: `http://genomebiology.com/2004/5/10/R80`.

[40] P. Giles and D. Kipling. “Normality of oligonucleotide microarray data and implications for parametric statistical analyses”. In: *Bioinformatics* 19 (2003), pp. 2254–2262.

[41] R. Gottardo et al. “Bayesian Robust Inference for Differential Gene Expression in Microarrays with Multiple Samples”. In: *Biometrics* 62 (2006), pp. 10–18.

[42] R. Gottardo et al. “Statistical analysis of microarray data: a Bayesian approach”. In: *Biostatistics* 4.4 (2003), pp. 597–620.

[43] P. Green. “Reversible Jump Markov chain Monte Carlo computation and Bayesian model determination”. In: *Biometrika* 82.4 (1995), pp. 711–732.

[44] B. Grün and F. Leisch. “Dealing with label switching in mixture models under genuine multimodality”. In: *Journal of Multivariate Analysis* 100.5 (2009), 851–861.

[45] J. de Haan et al. “Robust ANOVA for microarray data”. In: *Chemometrics and Intelligent Laboratory Systems* 98 (2009), pp. 38–44.

[46] J. Hardin and J. Wilson. “A note on oligonucleotide expression values not being normally distributed”. In: *Biostatistics* 10.3 (2009), pp. 446–450.

[47] B. Hill. “Inference about variance components in the one-way model”. In: *Journal of the American Statistical Association* 60 (1965), pp. 806–825.

[48] Huber et al. “Parameter estimation for the calibration and variance stabilization of microarray data”. In: *Statistical Applications in Genetics and Molecular Biology* 2(1).1 (2003).

[49] W. Huber, A. Heydebreck, and M. Vingron. “Error models for microarray intensities”. In: *Bioconductor Project Working Papers* (2004).

[50] W. Huber et al. “Variance stabilization applied to microarray data calibration and to the quantification of differential expression”. In: *Bioinformatics* 18 (2002), pp. 96–104.

[51] J. Ibrahim, M-H. Chen, and R. Gray. “Bayesian models for gene expression with DNA microarray data”. In: *J. Am. Stat. Assoc.* 97 (2002), pp. 88–99.

[52] R. Irizarry et al. “Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data”. In: *Biostatistics* 31 (2003), pp. 249–264.

[53] R. Irizarry et al. “Summaries of Affymetrix GeneChip probe level data”. In: *Bioinformatics* 31 (2003), e15.

[54] H. Ishwaran and J. Rao. “Detecting Differentially Expressed Gene in Microarrays using Bayesian Model Selection”. In: *J. Am. Stat. Assoc.* 98 (2003), pp. 438–455.

[55] J. Jin et al. “Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays”. In: *Journal of Pharmalcology and experimental therapeutics* 307.1 (2003), pp. 93–109.

[56] M. Kendall and J. Keith. *Time Series*. Arnold, 1990.

[57] M. Lee et al. “Nonparametric methods for microarray data based on exchangeability and borrowed power”. In: *Journal of Biopharmaceutical Statistics* 15 (2005), pp. 783–797.

[58] A. Lewin, N. Bochkina, and S. Richardson. “Fully Bayesian Mixture Model for Differential Gene Expression: Simulations and Model Checks”. In: *Statistical Applications in Genetics and Molecular Biology* 6 (2007).

[59] S. Li et al. “Assessment of diet-induced obese rats as an obesity model by comparative functional genomics”. In: *Obesity (Silver Spring)* 16.4 (2008), pp. 811–818.

[60] S. Li et al. “Assessment of diet-induced obese rats as an obesity model by comparative functional genomics”. In: *Obesity (Silver Spring)* 16.4 (2008), pp. 811–818.

[61] X. Liu, N. Milo M.and Lawrence, and M. Rattray. “A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips”. In: *Bioinformatics* 21(18) (2005), pp. 3637–3644.

[62] X. Liu, N. Milo M.and Lawrence, and M. Rattray. “Probe-level measurement error improves accuracy in detecting differential gene expression”. In: *Bioinformatics* 22(17) (2006), pp. 2107–2113.

[63] N. MacLennan et al. “Targeted disruption of glycerol kinase gene in mice: expression analysis in liver shows alterations in network partners related to glycerol kinase activity”. In: *Human Molecular Genetics* 15.3 (2006), pp. 405–15.

[64] G. Mclachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics, 2000.

[65] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer, 1996.

[66] F. Middleton et al. “Application of genomic technologies: DNA microarrays and metabolic profiling of obesity in the hypothalamus and in subcutaneous fat”. In: *Nutrition* 20 (2004), pp. 14–25.

[67] U. Mückstein et al. “Hybridization thermodynamics of NimbleGen Microarrays”. In: *BMC Bioinformatics* 11.35 (2012).

[68] J. Novak et al. “Generalization of DNA microarray dispersion properties: microarray equivalent of t-distribution”. In: *Biology Direct* (2006), doi: 10.1186/1745–6150–1–27.

[69] J. Novak et al. “Generalization of DNA microarray dispersion properties: microarray equivalent of t-distribution”. In: *Biology Direct* 1.27 (2006). DOI: `10.1186/1745-6150-1-27`.

[70] T. Park and D. van Dyk. “Partially Collapsed Gibbs Samplers: Illustrations and Applications”. In: *Journal of Computational and Graphical Statistics* 18.2 (2009), pp. 283–305.

[71] K. Pearson. “Contributions to the mathematical theory of evolution”. In: *Philosophical transactions of the Royal Society London, A* 185 (1894), pp. 71–110.

[72] M. Plummer et al. “CODA: Convergence Diagnosis and Output Analysis for MCMC”. In: *R News* 6.1 (2006), pp. 7–11.

[73] Martyn Plummer et al. “CODA: Convergence Diagnosis and Output Analysis for MCMC”. In: *R News* 6.1 (2006), pp. 7–11. URL: `http://CRAN.R-project.org/doc/Rnews/`.

[74] A. Posekany. “Robustness Issues in Bayesian Analysis of Microarray Data”. MA thesis. Technical University, Vienna, 2009.

[75] A. Posekany, K. Felsenstein, and P. Sykacek. “Biological Assessment of robust noise models in microarray data analysis”. In: *Bioinformatics* 27.6 (2011), pp. 807–814.

[76] E. Purdom and S. Holmes. “Error Distribution for Gene Expression Data”. In: *Statisitical Applications in Genetics and Molecular Biology* 4 (2005), Article16.

[77] R Development Core Team. *R: A Language and Environment for Statistical Computing.* ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2011. URL: `http://www.R-project.org/`.

[78] S. Richardson and P. Green. “On Bayesian Analysis of Mixtures with an Unknown Number of Components”. In: *Journal of the Royal Statistical Society, Series B* 59.4 (1997), pp. 731–792.

[79] C. Robert and G. Casella. *Introducing Monte Carlo Methods in R.* Springer-Verlag, 2009.

[80] C. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer-Verlag, 1999.

[81] Christian P. Robert. *The Bayesian choice.* Springer, 2001.

[82] G. Roberts and J. Rosenthal. "Harris Recurrence of Metropolis-Within-Gibbs and Trans-Dimensional Markov chains". In: *Ann. Appl. Prob.* 16 (2006), pp. 2123–2139.

[83] G. Roberts and J. Rosenthal. "Markov chain Monte Carlo: Some practical applications of theoretical results". In: *Can. J. Stat.* 26 (1998), pp. 5–31.

[84] G. Roberts and J. Rosenthal. "Two Convergence Properties of Hybrid Samplers". In: *The Annals of Applied Probability* 8.2 (1998), pp. 397–407.

[85] Gareth O. Roberts and Sujit K. Sahu. "Approximate Predetermined Convergence Properties of the Gibbs Sampler". In: *Journal of Computational and Graphical Statistics* 10.2 (2001), pp. 216–229. ISSN: 10618600. URL: http://www.jstor.org/stable/1391009.

[86] F. Ruggeri. "Nonparametric Bayesian robustness". In: *Chilean Journal of Statistics* 1.2 (2010), pp. 51–68.

[87] F. Schmid and M. Trede. "Simple tests for peakedness, fat tails and leptokurtosis based on quantiles". In: *Computational Statistics & Data Analysis* 43 (2003), pp. 1–12.

[88] B. Shahbaba and R. M. Neal. "Gene function classification using Bayesian models with hierarchy-based priors". In: *BMC Bioinformatics* 7 (2006), p. 448.

[89] N. D. Shyamalkumar. "Likelihood robustness". In: *In Robust Bayesian Analysis.* Ed. by David Rios Insua and Fabrizio Ruggeri. Springer, 2000.

[90] C. Small et al. "Profiling gene expression during the differentiation and development of the murine embryonic gonad". In: *Biol. Reprod.* 72.2 (2005), pp. 492–501.

[91] Gordon K. Smyth. “Limma: linear models for microarray data”. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Ed. by R. Gentleman et al. New York: Springer, 2005, pp. 397–420.

[92] M. Somel et al. “Human and chimpanzee gene expression differences replicated in mice fed different diets”. In: *PLoS One* 3.1 (2008), e1504.

[93] S. Someya et al. “The role of mtDNA mutations in the pathogenesis of age-related hearing loss in mice carrying a mutator DNA polymerase gamma”. In: *Neurobiological Aging* 29.7 (2008), pp. 1080–92.

[94] T. Speed, ed. *Statistical analysis of gene expression microarray data*. Chapman & Hall, 2003.

[95] M. Stephens. “Bayesian Analysis of mixture models with an unknown number of components: an alternative to Reversible Jump Methods”. In: *The Annals of Statistics* 28.1 (2000), pp. 40–74.

[96] M. Stephens. “Bayesian Methods for Mixtures of Normal Distributions”. PhD thesis. Magdalen College, Oxford, 1997.

[97] M. Stephens. “Dealing with label switching in mixture models”. In: *J. R. Stat. Soc. B* 62 (2000), pp. 795–809.

[98] D. Talantov et al. “Novel genes associated with malignant melanoma but not benign melanocytic lesions”. In: *Clin. Cancer Res.* 11.20 (2005), pp. 7234–42.

[99] D. Talantov et al. “Novel genes associated with malignant melanoma but not benign melanocytic lesions”. In: *Clin. Cancer Res.* 11.20 (2005), pp. 7234–7242.

[100] Tian et al. “The Role of the Wnt-Signaling Antagonist DKK1 in the Development of Osteolytic Lesions in Multiple Myeloma”. In: *N. Engl. J. Med.* 349.26 (2003), pp. 2483–2494.

[101] G. Tiao and W Tan. “Bayesian Analysis of random-effect models in the analysis of variance”. In: *Biometrika* 52 (1965), pp. 37–53.

[102] Luke Tierney. “Markov chains for exploring posterior distributions”. In: *The Annals of Applied Statistics* 22 (1994), pp. 1701–1762.

[103] V. Tusher, R. Tibshirani, and G. Chu. “Significance analysis of microarrays applied to the ionizing radiation response”. In: *Proceedings of the National Academy of Sciences* 98 (2001), pp. 5116–5121.

[104] G. J. G. Upton and A. P. Harrisson. “The Detection of Blur in Affymetrix GeneChips”. In: *Statistical Applications in Genetics and Molecular Biology* 9 (1 2010). Article 37. DOI: 10.2202/1544-6115.1590.

[105] D. Van Hoewyk et al. “Transcriptome analyses give insights into selenium-stress responses and selenium tolerance mechanisms in Arabidopsis”. In: *Physiol. Plant.* 132.2 (2008), pp. 236–53.

[106] R. Waagepetersen and D. Sorensen. “A tutorial on reversible jump MCMC with a view toward application in QTL-mapping”. In: *International Statistics Review* 69 (2001), pp. 49–61.

[107] P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman and Hall, 1991.

[108] K. Wang, S. Ng, and G. McLachlan. *Clustering of time course gene expression profiles using normal mixture models with AR(1) random effects.* 2011.

[109] Larry Wasserman. “The Conflict Between Improper Priors and Robustness”. In: *Journal of Statistical Planning and Inference* 52 (1996), pp. 1–15.

[110] E. Whitley and J. Ball. “Statistics review 6: Nonparametric methods”. In: *Critical Care* 6 (2002), pp. 509–513.

[111] Wikipedia. *DNA Microarray.* http://en.wikipedia.org/wiki/DNA$_{m}icroarray$.

[112] Y. Yang et al. “Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation”. In: *Nucleic Acid Research* 30 (2002), e15.

[113] Z. Yao et al. “A marfan syndrome gene expression phenotype in cultured skin fibroblasts”. In: *BMC Genomics* 8.39 (2007).

[114] Z. Yao et al. “A Marfan syndrome gene expression phenotype in cultured skin fibroblasts”. In: *BMC Genomics* 8:39 (2007).

[115] C. Yau and C. Holmes. “Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination”. In: *Bayesian Analysis* 6.2 (2011), pp. 329–352.

[116] H. Zhao et al. “Multivariate hierarchical Bayesian model for differential gene expression analysis in microarray experiments”. In: *BMC Bioinformatics* 9 (2008).

[117] J. Zimmerman et al. “Multiple mechanisms limit the duration of wakefulness in Drosophila brain”. In: *Physiol. Genomics* 27.3 (2006), pp. 337–50.