Master's Thesis

Robustness Issues in Bayesian Analysis of Microarray Data

carried out at the

Vienna Science Chair of Bioinformatics
Department of Biotechnology
University of Natural Resources and Applied Life Sciences, Vienna
and the Institute of
Statistics & Probability Theory
Technical University of Vienna

under the guidance of
Dr.techn. Peter Sykacek
and
Dr. techn. Klaus Felsenstein by

Alexandra Posekany
Sieveringerstr. 30A, 1190 Wien

Vienna, 9. 3. 2009

Abstract

In the past years many statistical methods and tools have been developed for the analysis of microarrays. Although it is a well-known problem that microarrays often produce widely dispersed data, little considerations about the robustification of the current methodology have been made. This work tests a possible approach of robustifying a hierarchical Bayesian ANOVA model, which is specifically designed for the analysis of microarrays, with respect to its underlying error model. Additionally, it means to provide an understanding of the differences of results compared to the standard model and their differing biological implications.

The core of the method is the model selection of a fitting likelihood function from a set of noncentral student's t distributions of different degrees of freedom and normal distributions. A hybrid MCMC sampler has been designed and implemented in Matlab in order to perform the model inference. It has been tested with several artificial and biological data sets.

Applying the method to different biological settings, has provided a clear answer to the question: is student's t distribution a more reasonable model distribution for such data sets? Student's t distributions with low degrees of freedom are generally preferred as error model. More importantly the results showed that differences between the robust (student's t) and the standard (Gaussian) model not only occurred in the statistical inference, but also led to different biological conclusions which were drawn based on Gene Ontology analysis. Thus this work shows the importance of handling the choice of model likelihood with great care in the field of microarray analysis.

Acknowledgements

It is a matter of 'mores maiorum' to read the roll of honour to all those who have supported the writer, which does not mean in the least that I want to say, they would not deserve such thanks for any other reason than carrying on long-established traditions.
I am forever grateful to my advisor Peter Sykacek for bearing my circulatory collapses and further escapades with admirable stoicism, and forcing me to overcome my minimalistic writing tendencies with unparalleled determination. It should not go unmentioned that I have profitted a lot from his knowledge and experience, if only I could have attained a tad of calm realism.
My special thank you goes to Klaus Felsenstein for being my knight in shining armour when I needed one throughout the process of getting the green light for this thesis. I am very grateful for all his well-meant advice and impulses, as well as the pedantic criticism only a mathematician is able to give (and to appreciate it).
Of course my thanks also go to all my colleagues of the VSC Bioinformatics who have supported me in various ways. Speaking of friends, I want to thank my friends Kara Ziehl for all the interesting talks and precious advice on writing style and Steffi Tauber for all the hours she has been listening to me snivelling about the horrendous proceedings of this works and for always providing me with a good meal to cheer me up, although I had to prepare it myself most of the times.
Above all I want to thank my family for supporting me despite not understanding a single word of this thesis and never having understood how someone in his/her right mind could even think of studying maths. As an emotional support especially my sister Kathi has been irreplaceable for me.

Whoever feels disregarded in the litany despite their self-sacrificing efforts, please forgive me I am just human including faults like forgetfulness and ignorance at times.

Contents

1 Introduction

1.1 Symbiosis of Biology and Statistics and The Motivation for this thesis

Ever since the days of Gregor Mendel statistical analysis of experimental results has been an indispensable part of biology especially genetics. Days have passed and the amount of data to be analysed has grown by several magnitudes but one basic notion has still remained the same: statistics is an irreplaceable tool for gaining results in an objective and well-founded manner. The enormous amount of data as well as the specialised design of experiments do not only call for computational toolboxes based on statistical methods they also require the design of statistical methods specifically focussing on the objectives of microarray experiments. The statistics language R [2] provides researchers with numerous packages developed and collected for such purposes by the Bioconductor project. Still the development and improvement of statistical methods is not finished yet and will not be so for a while, as novel measurement technology is demanding ever more realistic models of increasing complexity to deal with newly emerging biological questions.
Aside from the frequentistic methods, several probabilistic approaches have been made towards dealing with biological data. Such a fully Bayesian model has to be defined including the choice of underlying model and prior distributions. These choices are as much guided by experience or habit as the choice of classical statistical models is, who could imagine a regression model based on nonnormally distributed residuals after all? However such habits have to be checked and rechecked and new approaches have led to GLMs or other types of models. Bayesians are also all too familiar with the criticism of choosing model distributions mainly for reasons of convenience (as it may occur sometimes in case of conjugate priors) and a branch has grown focussing especially on this aspect of modelling, **robust Bayesian statistics** (see [11]). It focusses on 3 major aspects of any Bayesian model and the sensitivity of model outcomes on each of them: prior distribution, likelihood function and loss function (i.e. the inverse utility function [36]). It is the aim of this thesis to focus on the aspect of choosing a 'more robust' likelihood function, if it should be appropriate for the respective data.

1.2 Structure of the thesis

Chapter 2 provides a very brief introduction to the biological and technical methodology involved with microarrays. The same accounts for chapter 3 which gives an overview over a part of the field of Bayesian statistics and Bayesian robustness.

Chapter 4 presents the actual model and can be viewed as the mathematical backbone of this thesis. Furthermore it presents the actual focus of robustness in this thesis.

Chapter 5 is divided into two main parts. The first part provides an introduction to Markov chains and Markov Chain Monte Carlo sampling. The second part presents the sampling algorithm for the proposed robust regression model. This part can be seen as the methodological background for the actual implementation.

In chapter 6 we use artificial data sets for testing the algorithm and report inference results and convergence assessments.

In chapter 7 we finally demonstrate the application of the algorithm to some biological data sets and the conclusion from our investigation.

2 Introduction to Microarrays

The proposed model is specifically designed for the application to Microarray data. A short introduction shall be given for all those less familiar with the biological background. At first we focus a bit on nucleic acids, more specifically on RNA(Ribonucleic Acid) and DNA(Deoxyribonucleic Acid). Both are built by sequences of nucleotides, but while RNA is single stranded, DNA forms a double-helix as secondary structure. A nucleotide is a molecule consisting of phosphatase, one of the respective sugars Ribose (for RNA) and Deoxyribose (for DNA) and one of the 4 bases: Adenine(A), Guanine(G), Cytosine(C) and Thymine(T), which is replaced by Uracil(U) in case of RNA. When forming a double-helix only specific pairs of bases (G-C and A-T/U) can establish hydrogen bonds required for the stability of the helix.
This is an important fact for two essential cellular processes: ***transcription*** and ***reverse transcription***. Transcription is a natural process occurring in all cells. Information on the DNA is copied to a mRNA (messenger RNA) strand, which after certain chemical postprocessing steps will be used for the creation of proteins. The process of building a protein based on the information read from a mRNA is called ***translation***. Depending on the mRNA's half life few to several 100 copies of the same protein can be created from one mRNA. This straightforward process is the main notion behind the method of microarrays, which aims towards estimating the amount of mRNA, in order to estimate the amount of proteins. The process of ***reverse transcription*** does not occur naturally in any cell, unless in retroviruses. It is exactly the inverse process related to transcription as it writes the information of a mRNA strand back to DNA.

The term microarray refers to 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*** - on the solid surface, in a way that a quantitative measurement of a specific transcript of interest - the ***target*** - can be conducted. What makes the results of microarrays different from the ones of former biological techniques is that genome scale information of quantitative measurements is obtained.

Scientists differ between three types of microarrays regarding the analysed biological substance: ***tissue, protein and DNA***. ”Tissue microarrays immobilize small amounts of tissue from biopsies of multiple subjects on glass slides for immunohistochemical processing, while protein arrays immobilize peptides or intact proteins for detection by antibodies or other means”. [7]
The type of microarrays used in this thesis is the relatively young technology of DNA

microarrays, which even though they were first introduced in the 1990s, are most widely used. For the fabrication of microarrays a variety of technologies can be used, among them printing with fine-pointed pins onto glass slides, photolithography using pre-made masks, photolithography using dynamic micromirror devices, as well as ink-jet printing, or electrochemistry on microelectrode arrays.
Two main systems of DNA microarrays are commonly available:

- ***cDNA[1] microarrays***
  These are also referred to as ”spotted microarrays” , and are created by robotic spotting of PCR[2] products (primarily genes and expressed sequence tags). On cDNA chips two different types of targets respectively their interactions with the predefined probes can be observed. In order to be able to differ between the two labels with 2 different fluorescent dyes (usually a red-fluorescent dye, Cyanine 5 or Cy5, and a green-fluorescent dye, Cyanine 3 or Cy3) are attached to the DNA sequences. Based on the assumption of proportionality of brightness of the light of each frequency and the amount of DNA binding to the probes the brightness and thus this amount of DNA is measured. This raw data is the basic input for all further analyses.

- ***High-density oligonucleotide arrays***
  A trend towards these chips has occurred during the past years. Some of these, e.g. Affymetrix chips, are fabricated using photolithographic chemistry on silicon chips where light in combination with light-sensitive masking agents are used to create a sequence one nucleotide at a time across the entire array. The main difference between the results of these chips and the cDNA chip is the usage of just one type of cells and thus just one colour channel.

Last, but not least in this short introduction to microarrays, it should be mentioned that the objectives of microarray studies are crucial for determining the optimal data analysis. These can reach from the clear aim to distinguish between 2 (or sometimes more) groups of cells (males vs. females; patients vs. healthy control persons) to time courses of observations coming from one or more tissue, which are often used for inferring interactions among genes.

[1] complementary DNA, using the enzyme reverse transcriptase DNA, is synthesised from mRNA

[2] polymerase chain reaction: a molecular biological process which amplifies a DNA string generating millions of copies

3 Bayesian basics & Bayesian Robustness

3.1 The basics

Bayesian statistics is founded on the principle that any analysis is subjectified by its analyst and taking this into account in form of prior beliefs and prior distributions. This is represented in the principal theorem at the very heart of the theory.

Theorem 1 (Bayes' Formula).

Let H be an hypothesis, D the data, then

$$\mathbb{P}[H|D] = \frac{\mathbb{P}[D|H]\mathbb{P}[H]}{\mathbb{P}[D]} \tag{3.1}$$

In more generality this will become a statement about the uncertainty of a model parameter θ, which is represented by a *prior distribution* π on the parameter space Θ. While $f(x|\theta)$ represents the *likelihood function*. Then inference shall be based on the distribution of theta conditional under x, the *posterior distribution*,

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

The denominator is usually referred to as *marginal likelihood*, $m(x) = \int_\Theta \pi(\theta)f(x|\theta)d\theta$. As an introduction types of prior distributions will be presented, as some of these will play a role in the thesis. Before that another definition has to be presented, families of exponential distributions shall be formally defined and there will be a focus on their properties in some of the following considerations, since almost all distributions which are part of the statistical model presented in the thesis belong to such a family of distributions (see [36]).

Definition 1 (Exponential Family).

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

$$f(x|\theta) = C(\theta)h(x)e^{R(\theta)^T \cdot T(x)}. \tag{3.2}$$

In the special case if $R(\theta) = \theta$, i.e. $R(.)$ equals the identity $id_{\mathbb{R}^k}(.)$, the family is called natural.

Since the choice of a prior distribution and with this the way of getting prior information into a model is the key point of Bayesian analysis a short overview over types of prior distributions, their advantages and disadvantages shall be given.

- ***Natural conjugate prior***. The straightforward way of managing a Bayesian model is to choose a prior distribution with a structure similar to that of the likelihood function. This makes the prior better interpretable in terms of the model itself and makes adding prior information e.g. from equally structured earlier experiments easily manageable. From the point of view of calculation this has the added bonus of making the analytical determination of the posterior possible and thus updating fairly easy.

  When working with exponential type families the natural conjugate prior setting has the advantage of often describing data information in form of representative functions of the data, sufficient statistics. A more formal definition would be:

  Definition 2 (**Sufficient statistic**).

  When $x \sim f(x|\theta)$, a statistic $T(x)$ is said to be sufficient*, if the distribution of x conditional upon $T(x)$ is independent of θ, i.e. $p(x|T(x), \theta) = p(x|T(x))$.*

  A useful statement regarding sufficiency is made by the following factorisation theorem.

  Theorem 2 (**Factorisation theorem**).

  The density of x can be written in the form

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

  with g being a density of $T(x)$, iff T is sufficient.

  And the resulting theorem

  Theorem 3. *If a family of distributions $f(.|\theta)$ is such that for a sufficiently large sample size there exists a sufficient statistic of constant dimension, the family is exponential if the support of $f(.|\theta)$ is independent of θ.*

  The converse of this theorem is a natural property of any exponential family, i.e. such a sufficient statistic can be found for any exponential family. Thus a natural conjugate prior belonging to an exponential family itself can be found, it need only be compatible with the sufficient statistic. Because of their structure natural conjugate priors are of special interest for the simulation method of Gibbs sampling.

  However using conjugate prior always has the disadvantage that information is brought into the model, additionally only information that fits into 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 exact hyperparameters) is predefined by the model itself, this prior is also called *objective*, loosing some of subjectivity which for example an elicited prior (see [36]) would hold. The automation of the choice of a prior is as much an advantage as a nuisance. Thus easiness of computation should not be the only reason for choosing such a prior, its further advantages should still be carefully weighed against its disadvantages.

- ***Maximum Entropy prior***. The notion of this way of spinning prior information into a model is based on the information theoretic entity entropy which is a measure of uncertainty.

Definition 3 (Entropy).

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

Using the methodology of Maximum entropy it is possible to get information about certain characteristics of the prior into the model, as long as they can be written as prior expectations (e.g. moments, quantiles, . . .). A discrete prior which maximises the entropy (and thus the uncertainty), would then be formulated as

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

with λ_k being the Lagrange multipliers of the optimisation under the side conditions $\mathbb{E}_\pi[g_k(\theta)] = \omega_k$, where $\mathbb{E}_\pi$ describes the first moment of the distribution π of functions g_k of the parameter θ.
In the continuous case additionally a reference measure π_0 is required.

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

It is interesting to note that such a prior distribution will by definition be a member of an exponential family. A drawback of the method is that is often quite impractical and might produce impossible parameter values like negative variances ($g_1(\theta) = \theta, g_2(\theta) = \theta^2 \; \omega_1^2 > \omega_2$). Sometimes the moments used are incompatible and lead to a partial rejection of available information, e.g. contradictory definitions which force the analyst to drop one or more of the side conditions in order to obtain a density at all. (For details see [36])

- ***Noninformative prior***. One does not always have results of a prior experiment or as in explorative studies no information at all is available, yet it is still a necessity to convert this 'lack of information' into a prior distribution which is a necessary part of any Bayesian model. Since there is no ideal way of formulating a single function which represents complete ignorance, different types of 'non-informative' priors focus on different aspects of this lack of knowledge. One of these aspects is invariance to parameter transformations since one might use a transformation of the original parameter due to easier handling of the model, e.g. standard deviation or precision instead of variance. Yet when nothing is known about the original parameter, one is supposed to be in the dark about the transformed parameter as well.
  Along these lines a very general approach has been proposed by Jeffreys. Its notion is to base the calculation of the prior distribution on *Fisher's information matrix* (appropriate regularity conditions assumed)

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

The *Jeffreys noninformative prior distribution* would then be

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

which is invariant under diffeomorph parameter transformations.

A modification of Jeffrey's approach led to the notion of **reference priors** ([12]). The most notable difference between the methods is that the reference prior approach distinguishes between parameters of interests and nuisance parameters. A way to see the method in connection with the Jeffreys prior above has been presented by Robert ([36]). It is also a constructive method for obtaining the reference prior.

A model $x \sim f(x|\omega, \theta)$ depending on the multivariate parameter (θ, ω), where θ is the parameter of interest and ω the nuisance parameter. In order to obtain the reference prior one defines $\pi_J(\omega|\theta)$ as the Jeffreys prior of ω for fixed θ and calculates the marginal distribution

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

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

It is clear now that the reference prior is an extension of the notion of Jeffreys'. The reference prior equals the Jeffreys prior in cases where a normal approximation of the posterior is valid, which is the case for all continuous distributions as long as certain regularity conditions (see [12]) are fulfilled. In the discrete case the reference prior equals the uniform distribution.

For choosing the priors of the proposed model several aspects had to be taken into account. One of these aspects is the hierarchical structure of the model. Although it provides the statistician with some inherent amount of 'robustness' (see below), it requires a certain amount of manageability which is rarely a property of noninformative priors, but is a natural property of conjugate priors. Despite belonging to an exponential family in any case, the Maximum entropy prior need not be easy to handle. However the main reason for not using it is that for most microarray experiments not enough information is available a priori in order to specify any proper side conditions.

What turns the balance in favour of the natural conjugate prior for most cases is the usage of Gibbs sampling methods and simplification of updates for Metropolis-Hastings steps, which will be described in more detail in a separate chapter. However noninformative priors are chosen in several cases. For the discrete state space of the degrees of freedom parameter a uniform prior is optimal for manageability in updates as well as minimizing the information brought in by the prior, in some cases the noninformative priors can be even viewed as limiting distributions of the natural conjugate prior (Beta, Gamma priors).

3.2 Bayesian Robustness

The aim of Bayesian robustness is to smartly choose priors, likelihood or loss functions in such a way that they are less sensitive to changes of other model components. The basic idea of doing this is to define a class of distributions, which may work as priors or likelihoods, instead of choosing a single type of distribution for that purpose. The selection of natural conjugate or noninformative priors can be seen as an example for cases, when a single type of distribution is selected with a specific goal in mind, computational and interpretational practicality of the conjugate are weighed against compatibility with transformations of non-informative priors. However a problem even with so-called non-informative priors is that one single distribution cannot sufficiently express indifference about the parameter. A good statement in that respect has been made by Walley [47]:

> *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.*

A robustification of the situation might be to define a class which includes both the natural conjugate and non-informative and other types of prior distributions to cover a larger range of possible model behaviour.
There are 2 main problems that occur, when working with exponential family distributions (see [11]):

- exponential family distributions are very sensitive against outliers.
- conjugate priors have great influence in cases when the data jars with the prior information implicitly introduced by its specification, i.e. the informative choice of hyperparameters is even more influential if the data itself is not fully compatible with the parametric model structure.

3.2.1 Global Robustness

Because the concept of global robustness of priors will play a certain role in the focus of this work, the basic ideas of this theory will be presented. The principle behind it is that a class of prior distributions Γ is defined in such a way that it contains all "reasonable" distributions. The range of results determined from all models with priors in this class serves as an indicator whether the model is sufficiently robust: if the range $r(\Gamma)$ is not "too large" the results are considered as robust. (see [11]) The concept is rather vaguely defined and leaves it to the statistician to decide one the thresholds for "too large" as well as the quantity of interest.

$$\begin{array}{rcl} r(\Gamma) & = & \|\overline{\psi} - \underline{\psi}\|, \\ & & \overline{\psi} = \sup_{\pi\in\Gamma} \psi(\pi, f), \quad \underline{\psi} = \inf_{\pi\in\Gamma} \psi(\pi, f), \end{array} \tag{3.3}$$

where π represents the prior, f the likelihood function and $\psi(\pi, f)$ a point estimate from the posterior or another quantity of interest.

As the monotony criterion

$$\Gamma' \subseteq \Gamma \Rightarrow (\overline{\psi}' - \underline{\psi}') \leq (\overline{\psi} - \underline{\psi}) \tag{3.4}$$

holds, the range of results can be reduced by imposing reasonable restrictions on the class Γ and hereby gaining a subset Γ', which has a smaller range of results.

3.3 Likelihood robustness

In the majority of cases Bayesian robustness consideration focus on the robustification of prior distributions. There are 2 main reasons for this, firstly since the early days of Bayesian analysis the priors as subjective part of the method have been viewed as the weakest link of the theory thus being in the focus of most criticism. Yet logically the likelihood function has considerable influence, as it determines the way in which the data will influence the results. However there is no easy way of quantifying the actual influence, which leads us to the second reason why too many considerations of likelihood robustness have been avoided: investigation of the posterior robustness with respect to the likelihood is not an easy task.
Shyamalkumar ([42]) has proposed a method to investigate this, which works analogously to global robustness of priors (Berger's original concept is defined for priors and likelihood functions alike). Again a class of distributions Γ_f, from which the model likelihoods shall be chosen, is defined and the range of results shall give indication of how robust the model is (see equation (3.3)).

$$\overline{\psi} = \sup_{f \in \Gamma_f} \psi(\pi, f), \underline{\psi} = \inf_{f \in \Gamma_f} \psi(\pi, f), \tag{3.5}$$

Another way of investigating likelihood robustness is to choose 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 one looks for the 'optimal' model to determine the most robust behaviour.
The advantage of this method is that unlike the determination of global infima and suprema the complexity of calculation does not increase significantly with the increase of sample size. Its 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.

Although we could choose from a broad class of symmetric, 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 and normal distributions, in order to have models which are analytically tractable. This approach of robustification mainly focusses on outliers of the observations. The hierarchical structure of the proposed model makes sure a certain robustness with respect to the specification of priors is obtained.
An analysis of robustness with respect to the range of the posterior distribution or certain parameter estimates is virtually impossible since these quantities of interests are

determined by Markov chain Monte Carlo simulation. Thus more than one run per model has to be performed in order to reduce variation introduced by the simulation method itself and these combined results present 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. Thus Shyamalkumar's idea of finite classes 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 this thesis 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 microarrays often produce widely dispersed data. The commonly used models for determining gene expression are based on Gaussian distribution settings which provide analytically tractable results (e.g. see [28]). For example Baldi et al. [9] use t-tests with appropriate adjustments for the number of tests performed. Others have introduced fully Bayesian models based on normal distribution assumptions, as Ibrahim et al. [26], Zhao et al. [48] and Gottardo et al. [23]. All these approaches have in common that the high probability of 'extreme' values frequently appearing in microarray data is not suitably represented by the normal distribution model.

A statistical technique for determining the differential expression of genes, estimating and controlling error rates by the means of a non-parametric statistic has been introduced by Tusher et al. (see [46]). Using 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, but it is not the kind of robustness we are aiming for. We want to stay close the parametric model of normal distributions, but take into account data which deviates from the Gaussian distributions setting, e.g. far outlying data points. However, as we work with a linear regression model, we still want 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 been made, mainly focussing on Gaussian mixture distributions ([30]), rarely on t distributions ([24]). In some ways our modelling attempt is similar to Gottardo et al.'s ([24]), yet in others ours is more general. In contrast to the approach by Gottardo, we do not aim to compare the student's t approach against other methods of the types described above, but aim for directly comparing it to the same model in standard setting, i.e. Gaussian error distribution. We will elaborate in more detail, once our model has been described.

4 Mathematical Structure of the model

4.1 Model structure

The Bayesian hierarchical model which we will use in this thesis for investigating robustness is based on a latent variable implementation of a biological indicator variable I_g which furthermore is linked with an ANOVA type linear model:

$$y_{n,g} = x_{n,g}^T \beta_g + \varepsilon_{n,g}, \qquad n = 1, \dots, N, g = 1, \dots, G \tag{4.1}$$

where for any given sample n and gene g:

$y_{n,g}$ is the observed gene expression,

$x_{n,g}$ is a vector of the underlying design matrix;
$x_{n,g} = [\mathbb{I}(S_{n,g} = 1), \dots, \mathbb{I}(S_{n,g} = S)]^T \in \mathbb{R}^{S \times 1}$

$S_{n,g}$ is an factor variable encoding the biological system of observation $y_{n,g}$, known from the experimental design

β_g is the vector of means fitted by the model conditional on the indicator of (non-)differential expression

I_g is the biological indicator which differs between differential expression and no differential expression and thus determines the dimension of β_g

$\varepsilon_{n,g}$ noise residuals

The design matrix $(x_{1,g}, \dots, x_{N,g}) =: X_g = X \in \mathbb{R}^{S \times N}$ is of course independent of the gene g, as all genes have to appear in all experiments and systems. The actual parameter of interest in this setting is the biological indicator I_g, it will help to rank genes according to their posterior probability of being differentially expressed. In the model this indicator differs between between a univariate and a multivariate linear model by determining the dimension of β_g. A onedimensional parameter can be interpreted as the estimator of the mean for a gene for which we have equal means of intensities in all biological systems; this is the definition of ”no differential expression”.

$$\begin{aligned} I_g = 0: \quad \beta_{g,0} | I_g = 0 \quad &\sim \quad N_1(\mu_{g,0}, (\tau_{g,0})^{-1}) \\ &\beta_g = [\beta_{g,0}, \dots, \beta_{g,0}]^T \ \in \mathbb{R}^{S \times 1} \end{aligned} \tag{4.2}$$

However a multivariate vector β_g contains the different estimates of means for the respective groups, its dimension is naturally equal to the number of different groups. If the estimate of a group is distinguishable from at least one of the others, the gene g is called differentially expressed.

$$\begin{aligned} I_g = 1: \quad \beta_g | I_g = 1 \quad &\sim \quad N_S(\mu_g, T_g^{-1}) \\ &\mu_g = [\mu_{g,1}, \dots, \mu_{g,S}]^T \ \in \mathbb{R}^{S \times 1} \\ &T_g = \begin{pmatrix} \tau_{g,1} & & 0 \\ & \ddots & \\ 0 & & \tau_{g,S} \end{pmatrix} \in \mathbb{R}^{S \times S} \end{aligned} \tag{4.3}$$

Every gene has a certain probability of being differentially expressed, thus I_g itself will a priori be modelled by an alternative distribution with probability p of 'success', i.e. differential expression.

$$I_g | p \quad \sim \quad Bin(1, p) \tag{4.4}$$

For the update of the probability p a Beta distribution is chosen as prior for this parameter which is the natural conjugate prior.

$$p \quad \sim \quad Be(a, b). \tag{4.5}$$

This choice is justified not only by the easiness of updating, but especially by taking into account the 'counting' setting, which means that the total number of ones is the value of interest as well as a sufficient statistic in this model. Following the line of argumentation in the previous chapter this tips the scales in favour of the conjugate prior.

The part of the model we have not focussed on yet is the noise model which usually is a normal distribution for linear regression models. As an aim of this thesis is to provide robustness with respect to the assumed distribution of the observations, which after centralisation is equivalent to the distribution of the error term, the ansatz of the model will be to allow for a wide variety of possibilities for this error distribution. The selection of the most suitable error distribution can be handled a posteriori by model comparison or be a part of the algorithm as well.

4.2 Student's t Model

For the ansatz we take the general framework of the model described in section 4.1. Hereby the approach towards robustification is made using Student's t-distributions in the likelihood and prior setting. It is a well known fact which can be easily proved in general, that according to its definition the non-central t-distribution can be replaced by an hierarchical structure consisting of a Normal- and a Gamma-distribution in the following way:

$$X \sim t_\nu(\mu, \sigma^2) \Leftrightarrow \begin{array}{l} X|\varphi \sim N(\mu, \frac{1}{\varphi}\sigma^2) \\ \varphi \sim Ga(\frac{\nu}{2}, \frac{\nu}{2}) \end{array} \tag{4.6}$$

According to (4.6) the model is written as

$$\begin{array}{rcl} y_{n,g}|\beta_g, \nu & \sim & t_\nu(x_{n,g}^T\beta_g, \tau_\varepsilon)^{-1}) \;\; \Leftrightarrow \\ y_{n,g}|\beta_g, \varphi & \sim & N(x_{n,g}^T\beta_g, (\varphi_{n,g}\tau_\varepsilon)^{-1}) \\ \varphi_{n,g}|\nu & \sim & Ga(\frac{\nu}{2}, \frac{\nu}{2}) \\ \tau_\varepsilon|g, h & \sim & Ga(g, h) \end{array} \tag{4.7}$$

The auxiliary parameter $\varphi_{n,g}$ can be interpreted as a scaling factor which rescales the variance of the normal distribution such that outlying values become more probable. This is the robust behaviour of t distribution which we want to gain for our noise model.

A necessary condition for this model to work is to show that the marginal distribution of $y_{n,g}$ is indeed a student's t distribution. Although the interrelation between student's t distribution and normal distribution is well-known, it will be proved anyway for reasons of completeness.

Lemma 1. *The marginal distribution $m(y_{n,g}|\nu)$ of $y_{n,g}$ is t-distributed with degrees of freedom ν.*

Proof.

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

The structure of the expression $I(\varphi_{n,g})$ above is the same as a Gamma-distribution $Ga(a, b)$ with parameters for shape $a = \frac{\nu+1}{2}$ and rate $b = \frac{1}{2}(\nu + \tau(y_{n,g} - x_{n,g}^T\beta_g)^2)$ except for the normalisation constant. Thus the marginal distribution equals

$$m(y_{n,g}) = \int_0^\infty I(\varphi_{n,g})d\varphi_{n,g} = c_1 c_2 \frac{\Gamma(a)}{b^a}$$

which after cancelling a few terms results in

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

□

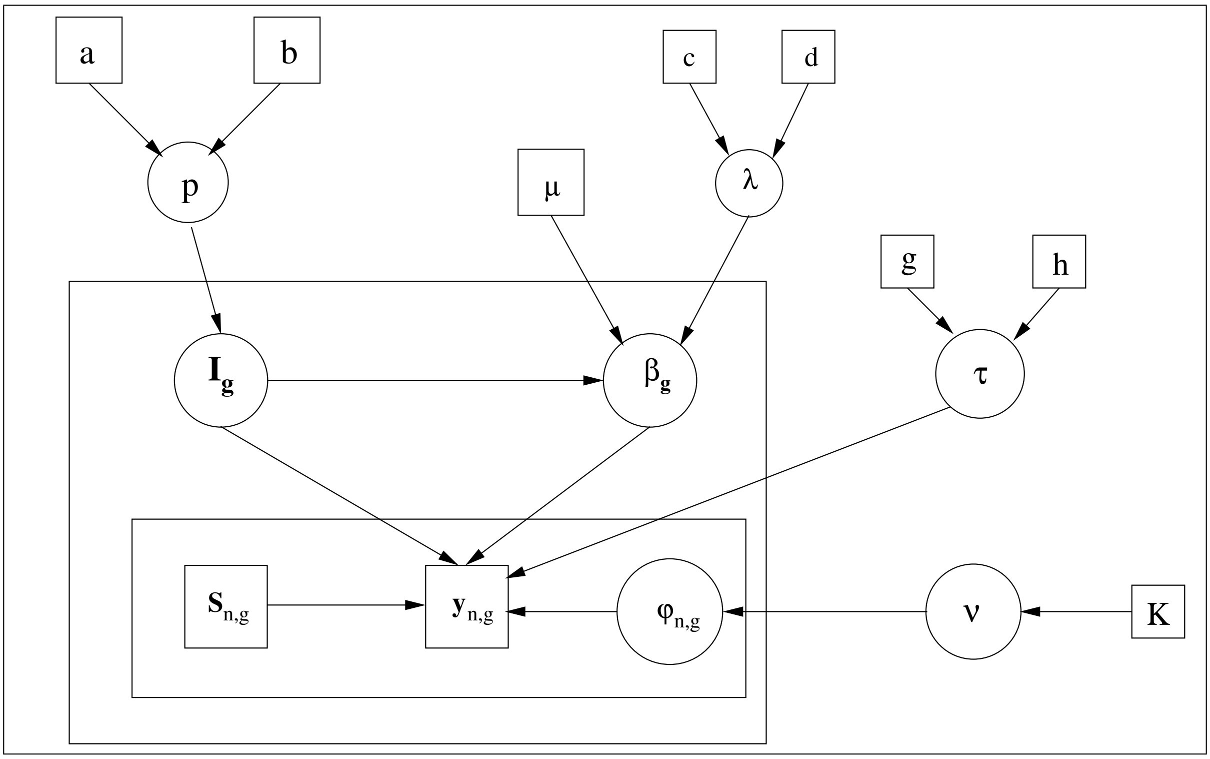


Figure 4.1: Directed Acyclic Graph representation of the model rectangular frames refer to variables which are fixed during the updates (data, fixed hyperparameters), variables in circles are updated as parts of the model

$y_{n,g}$	observations, i.e. normalised light intensities
$S_{n,g}$	indicator to which experiment type observation $y_{n,g}$ belongs
β_g	ANOVA parameter vector for gene g
I_g	indicator of differential expression
p	probability of a gene to be differentially expressed
λ	prior precision of β_g
τ	precision of the regression model
$\varphi_{n,g}$	scaling parameter linking normal and t distribution
ν	degrees of freedom of the error model

An essential component of the model in figure 4.1 is a student's t noise model of varying degrees of freedom. This model is set up such as to allow us to consider robustness issues with respect to outliers in the data $y_{n,g}$. ν decodes the degrees of freedom of a t distribution, thus for high enough values the t distributions will be sufficiently similar to normal distributions that differing between them does not make any sense. Therefore a cut-off value ν_{max} is specified for determining the value where normality can be assumed, i.e. reaching the maximum value is equivalent with choosing a normal distribution model. However this model is not approximated by the $t_{\nu_{max}}$ distribution, but an exact normal distribution model is used. In order to implement such a setting moving between parameter spaces of different dimension is required and will be realised by a reversible jump move within the MCMC algorithm.
In order to gain flexibility with respect to the choice of degrees of freedom for the t-distribution a discrete uniform hyperprior on the set $\mathfrak{N}$ over the parameter ν is specified:

$$\begin{aligned} \nu &\sim U_{\mathfrak{N}} && (4.8)\\ & \mathfrak{N} := \{x \in \mathbb{R} | 1 \leq x := j \cdot c_{grid} \leq \nu_{max}, j \in \mathbb{N}\} && (4.9)\\ \Leftrightarrow\ & \mathbb{P}[\nu = k|K] = 1/K, \ \ k \in \mathfrak{N}; \ \ K = |\mathfrak{N}| && (4.10) \end{aligned}$$

The choice of a uniform prior on this finite set also represents our lack of information regarding the underlying noise model. In order to improve readability the 'size' with respect to the counting measure of the set $\mathfrak{N}$, K, is used for the specification of the uniform distribution in figure 4.1.
However the definition of the set (see equation (4.9)) offers us great flexibility in the choice of the underlying parameter space and thus the analysis of robust behaviour. Choosing a grid size c_{grid} equal to 1 or even 5 allows us to work with clearly distinguishable t distributions, whereas refining the grid allows us to approximate 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.

The integration of the degrees of freedom parameter into the model makes it possible to let the model choose itself which error distribution is the most suitable. It allows us to take such a large number of models into account The biological indicator for differential expression follows a Bernoulli distribution

$$\pi(I_g|p) \ = \ p^{I_g}(1-p)^{1-I_g} \qquad (4.11)$$

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

$$\pi(p) \ = \ \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}. \qquad (4.12)$$

Conditional on "differential expression" respectively "no differential expression", the coefficient vector is determined by a multidimensional respectively one-dimensional underlying distribution as described above in section 4.1.
As a special case of the general settings above several restrictions for the parameters

involved are made. The hyperparameter μ is assumed to be fixed, i.e. $\mu_{g,s} = \mu \ \forall g, s$, e.g. taking the value of the overall sample mean, whereas the precision of β_g shall be specified by the parameter λ, which shall be common parameter for all prior precision parameters and follows a Gamma distribution, i.e.

$$\begin{aligned} \tau_{g,s} &:= \lambda \ \ \forall g, s && (4.13) \\ \lambda &\sim Ga(c, d) && (4.14) \end{aligned}$$

This reduces the model parts (4.2), (4.3) to:

$$\begin{array}{lll} I_g = 0 \quad \beta_{g,0}|I_g & \sim & N_1(\mu, (\lambda)^{-1}) \\ & & \beta_g = [\beta_{g,0}, \ldots, \beta_{g,0}]^T \ \in \mathbb{R}^{S\times 1} \\ I_g = 1 \quad \beta_g|I_g & \sim & N_S(\mu, (\lambda)^{-1}E_S) \end{array} \qquad (4.15)$$

The following table gives an overview over the model parameters and their distributions:

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

Table 4.1: Overview over Student‘s t model

Robustness in this thesis

As already mentioned above, different components of a probabilistic model can be aims for robustness considerations. The main focus of this work however is the robustification of the likelihood function of a hierarchical ANOVA model. The standard distribution setting for such a model would be a Gaussian error distribution (see [26], hierarchical model [48]; Bayesian ANOVA for microarrays [28]). Using Student's t distributions in order to gain robustness compared to a Gaussian distribution based model has been proposed several times, among others by Berger ([11]). In the context of microarrays it has been used by Gottardo et al. ([24]). The fact that the student's t distribution has higher probability on its tails makes it a reasonable candidate for models wishing to take outlying values into account. At the same time it shares certain properties with the normal distribution, like symmetry and unimodality. These properties are important for residuals of a regression model. Thus the student's t distribution is applicable for modelling values which behave like Gaussian values except for a higher probability of 'outlyingness'. Since we are working in the framework of ANOVA it is only necessary to take care of outliers in the observations $y_{n,g}$. This is also a good reason why this approached is focussed mainly on robustification of the likelihood function, which is linked to the behaviour of the observations.

To show the ansatz of the robustification in the framework of Bayesian Robustness studies as performed by Jim Berger, 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:

$$\Gamma = \{\{t_\nu(\mu, \tau^{-1}), \nu \in \mathfrak{N} \setminus \{\nu_{max}\}\}, N(\mu, \tau^{-1})\} \tag{4.16}$$

The definition of the set $\mathfrak{N}$ in (4.9) makes this approach very flexible. Choosing only a few values for ν allows us to make clear decisions of the tendency towards normality respectively 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. This discretisation is of special importance for the 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 it is supposed to approximate. Thus an upper bound for ν is important in order to make a clear decision when the distribution is sufficiently similar to a normal distribution to no longer have need of robustification w.r.t outliers. 'Jumping' to a normal distribution model, when this 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 structure of the presented model, mainly the variable dimensions of $\beta_g|I_g$, makes finding an analytical 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 important. 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). It will rather be the purpose of the model to indicate, whether there exists a problem in principle with the assumption of normally distributed data, on which further analysis steps would be based. The variable degrees of freedom parameter ν is supposed to give an answer to that question.

As mentioned above, our model is in some ways comparable to the approach by Gottardo et al. ([24]). However our goals differ, as we aim for comparing the model to its normal distribution analogue, in order to answer the questions, if a student's t model is required at all and how "far away" from a normal distribution in terms of degrees of freedom we truly are. Additionally we have defined the set of t distributions to include in a more general and flexible way. Firstly we can differ between t distributions with clearly different degrees of freedom values which is useful in principle, but might be problematic in other respects. Secondly we can reduce the step size far enough such that ν can be viewed as discretisation of a continuous degrees of freedom parameter, while at the same time we keep the advantages of the discrete setting described above. Using test data sets we will show the advantage of using a smaller step size in addition to the larger one. Additionally the variable dimension of $\beta_g|I_g$ makes a big difference between their ansatz and our model.

5 MCMC schemes

5.1 Markov Chain Monte Carlo Methods

As its name is telling already MCMC methods are based on 2 concepts of mathematics respectively computational integration:

1. Markov Chains
2. Monte Carlo integration

MCMC uses the approximation of expectations by means of random draws from a given distribution in combination with Markov chains which under certain conditions behave like draws from a single stationary distribution. The most important background of both theories will be presented in a short and compact way, as far as it is relevant for this work. The detailed treatment of the following definitions and theorems is a requirement for analysing the theoretical behaviour of the algorithm, as far as it can be deduced from the transition kernels of its components. As this is nontrivial, it is necessary to have a detailed understanding for the terms which are building the concept. Especially the chapter about Markov chains will contain a lot of terms which have no obvious use for the algorithm, but they create a basis for the following terms and statements.

5.1.1 Monte Carlo integration

The generic problem for classical Monte Carlo integration is the calculation of the following term:

$$\mathbb{E}_f[h(X)] \quad = \quad \int_{\mathcal{X}} h(x)f(x)dx \tag{5.1}$$

Given the observations $(X_1, \ldots, X_n)$ which have been generated from the density $f(.)$ the expression (5.1) can be approximated by the *empirical average*

$$\overline{h_n} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} h(x_i) \tag{5.2}$$

Due to the Strong Law of Large Numbers $\overline{h_n}$ converges almost surely to $\mathbb{E}_f[h(X)]$. Especially under the condition that h^2 has finite expectation under f, the speed of convergence can actually be assessed, which will be of importance for the *construction of convergence tests* for the method. This follows as the variance of h_n is

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

and its empirical estimator is

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

Thus the term

$$\frac{\overline{h_n} - \mathbb{E}_f[h(X)]}{\sqrt{v_n}} \quad \dot{\sim} \quad N(0,1)$$

5.1.2 Markov Chain theory

Definition 1 (**Markov Chain**).

A Markov chain *is a sequence of random variables $X_1, X_2, \ldots$, which fulfills the* Markov property*,*

$$\mathbb{P}[X_{n+1} = x | X_n = x_n, \ldots, X_0 = x_0] = \mathbb{P}[X_{n+1} = x | X_n = x_n], \tag{5.3}$$

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

An alternative way of defining Markov chains is via the term of its transition kernel, which is the function that determines the transition between the chains 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. *$\forall x \in \mathcal{X} : \mathcal{K}(x, .)$ is a probability measure, i.e. for every fixed value of the state space $\mathcal{X}$ the function $\mathcal{K}(x, .)$ operates on the Borel sets and assigns a probability (depending on x) to every set of $\mathcal{B}(\mathcal{X})$;*
2. *$\forall A \in \mathcal{B}(\mathcal{X}) : \mathcal{K}(., A)$ is measurable, i.e. for every fixed set A the function $\mathcal{K}(., A)$ operates on the state space $\mathcal{X}$ and is measurable.*

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

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

In the general case, the probability of reaching a set A when starting from x is

$$\mathbb{P}_x(X_1 \in A) = K(x, A).$$

A *Markov chain* is then defined as sequence of random variables which fulfills the Markov property in (5.3) and allows expressing the probability to reach a point in the set A when coming from $X_n = x_n$ as

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

The chain is *homogeneous*, if the distribution of $(X_{t_1}, \dots, X_{t_k})|x_{t_0}$ is the same as the distribution of $(X_{t_1-t_0}, \dots, X_{t_k-t_0})|x_0$.

An important property of transition kernels is expressed 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 $(m,n) \in \mathbb{N}^2, x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})$,

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

These equations describe that the probability to reach a set A in $m+n$ steps when starting from x by taking into account all interim values which can be reached from x and allow us to reach A within a certain finite number of steps each. Such a principle will be important for the notion of irreducibility. Another important term when dealing with Markov chains is the following:

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 $X_n \notin A\ \forall n$.

Also associated with a set A is the number of passages *of (X_n) in A,*

$$\eta_A = \sum_{n=1}^{\infty} \mathbb{I}_A(X_n)$$

Related to this term is the probability of return to A in a finite number of steps,
$\mathbb{P}[\tau_A < \infty]$.

There are several properties to look at, when studying a Markov chain's sensitivity to its initial conditions. Among the first is irreducibility. The possibility of reaching any point in the state space in a finite number of steps independently of where the chain starts is described by this notion.

Definition 4 (Irreducibility).

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

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

Given an auxiliary measure μ, the Markov chain (X_n) with transition kernel $\mathcal{K}(x,y)$ is μ-irreducible *if, for every $A \in \mathcal{B}(\mathcal{X})$ with $\mu(A) > 0$, there exists an n such that*

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

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

Certain properties are sufficient in order two imply irreducibility of a chain.

Theorem 2 (Irreducibility of (X_n)).

The 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:

- $\exists n \in \mathbb{N} : \mathcal{K}^n(x, A) > 0$
- $\mathbb{E}[\eta_A] > 0$*;*
- $\mathcal{K}_\varepsilon(x, A) = (1 - \varepsilon) \sum_{i=0}^{\infty} \varepsilon^i \mathcal{K}^i(x, A) > 0$ *for an ε with $0 < \varepsilon < 1$*

Among all probability measures with respect to which a chain is irreducible, one is of special interest, 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 ([37]).

A definition of high theoretical relevance is the definition of atoms and small sets.

Definition 5 (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 accessible *if $\psi(\alpha) > 0$.*

By its 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. It requires that such a 'minimum' probability of reaching a set A exists for a positive number of steps.

Definition 6 (Small Set).

A set C is small *if there exist $m \in \mathbb{N}^*$ 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 for which a guarantee exists that any set $A \in \mathcal{B}(\mathcal{X})$ can be reached in a given number of steps m with a positive probability which is 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 any Borel set A can be reached with positive probability. We set $\mathcal{K}^N(x, A) =: \nu_N(A)$ for the maximum of these n, N, which exists, as we only have a finite number of states in X. Due to the Chapman-Kolmogorow equations 1 this gives us 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 7 (Periodicity).

A state x has period d if a return to state x must occur in multiples of d time steps, i.e.

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

(gcd denotes the greatest common divisor).
If the chain is irreducible, which implies 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 freedom of moving 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. This leads us to the property of recurrence, which can be viewed in a discrete setting as a 'guarantee of a sure return'. Of course it is satisfied for any irreducible chain on a finite space.

Definition 8 (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, $\mathbb{E}_x[\eta_x]$, is finite. If this is not he case, i.e. $\mathbb{E}_x[\eta_x] = \infty$, it is* recurrent*.*

These properties of single states apply to the whole chain, if the chain if irreducible, which follows from the Chapman-Kolmogorow equations. Furthermore for any Markov chain the following definition applies

Definition 9 (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, $\psi(A) > 0 :\ \mathbb{E}_x[\eta_A] = \infty\ \forall x \in A$*

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 the following classification result that recurrence and transience are dichotomous properties for ψ-irreducible chains holds:

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

A ψ-irreducible chain is either recurrent or transient.

A more rigid property, which requires not only an infinite average number of visits for every small set - which implies the same limiting behaviour of the chain for almost every starting value - but applies to all states, is Harris recurrence.

Definition 10 (Harris recurrence).

A set A is Harris recurrent *if* $\mathbb{P}_x[\eta_A = \infty] = 1 \ \ \forall x \in A$.
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.

Theorem 4 (Harris recurrence of (X_n)).

If for every $A \in \mathcal{B}(\mathcal{X})$, $\mathbb{P}_x[\tau_A < \infty] = 1 \ \ \forall x \in A$, then $\mathbb{P}_x[\eta_A = \infty] = 1 \ \ \forall x \in \mathcal{X}$, and (X_n) is Harris recurrent.

Proof. see [37], 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
$\mathbb{P}_x[\tau_C < \infty] = 1 \ \ \forall x \in \mathcal{X}$, then (X_n) is Harris recurrent.

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.
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 means 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 11 (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 *- recurrence is fulfilled automatically by an irreducible chain. 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 shall clarify the connection between positivity and recurrence.

Theorem 6 (Positive recurrence).

If the chain X_n is positive, it is recurrent.

Kac's theorem is a rather classical result on irreducible Markov chains in a discrete state-space. In principle it states that when the stationary distribution exists, it is given by

$$\pi_x = (\mathbb{E}_x[\tau_x])^{-1}$$

An implication of this result is that $(\mathbb{E}_x[\tau_x])^{-1}$ is the eigenvector associated with the eigenvalue 1 of the corresponding transition matrix. This result can also be generalised for the continuous case. An implication of that 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.

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 in principle states that the dynamics of the chain is not influenced by the direction of time. More formally this means

Definition 12 (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 13 (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 K (and its respective Markov chain). A more general statement links this condition with the notion of reversibility.

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

Suppose that a Markov chain with transition kernel $\mathcal{K}$ satisfies the detailed balance condition with π a probability density function. Then:

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. □

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

Definition 14 (**Ergodicity**).

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

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

For the next 2 definitions the following norm of a measure μ is used:

$$\|\mu\| = sup_{|g|\le 1}\left|\int g(x)\mu(dx)\right| \tag{5.4}$$

A chain is geometrically ergodic*, if for a real-valued function M with $\mathbb{E}_\pi[|M|] < \infty$ and $0 < r < 1$*

$$\|K^n(x,.) - \pi\| \le M(x)\cdot r^n$$

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

$$\sup_x \|K^n(x,.) - \pi\| \le M \cdot r^n$$

Definition 15 (**Total Variation norm**).

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

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

Among the many statements that can be made about convergence under these conditions the most important are

Theorem 9 (Convergence in the Total Variation norm).

If (X_n) is Harris positive and aperiodic, then

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

for every initial distribution μ.

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

Theorem 10 (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.*

This section means to show how certain properties of the Markov chain lead to conclusions about its behaviour. 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. Practical implications are unfortunately very limited,as all these notions ranging from irreducibility to convergence assume an infinite number of draws.

5.2 Overview over some important sampling methods

The principle of any MCMC method is that if one cannot sample from the distribution ξ directly to use an ergodic Markov chain with stationary distribution ξ for obtaining samples of just that distribution ξ. The most universal of sampling schemes is the Metropolis-Hastings sampler.

5.2.1 Metropolis Hastings Sampler

The aim of the Metropolis Hastings sampler is drawing from the objective ***target density*** ξ which will be realised via an auxiliary conditional distribution $q(.|.)$ of a proposed value given the 'old' value, the ***proposal density***, which 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:

- 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 $\quad Y_t \quad$ with probability $\alpha(x^{(t-1)}, Y_t)$ or

     stay at the old value $\quad x^{(t-1)} \quad$ with probability $1 - \alpha(x^{(t-1)}, Y_t)$

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

     The transition kernel of the Metropolis-Hastings sampler is

$$\mathcal{K}(x, y) = \alpha(x, y)q(y|x) + (1 - \int \alpha(x, y)q(y|x)dy)\delta_x(y) \tag{5.5}$$

Table 5.1: Generic Metropolis-Hastings sampling algorithm

If the ratio of target and proposal function in table (5.1) is increased for the proposal compared to the old value, then the value is accepted for sure. Otherwise, if the ratio decreases, the proposal is accepted with probability α.

The approach may be illustrated by a simple example.

Example 1. *Metropolis-Hastings Sampler for student's t distribution*

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

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

To ease the situation, 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:

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

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

Certain conditions are required for the functions which define the Metropolis-Hastings acceptance probability and transition kernel in order to draw conclusions about properties of the chains necessary for convergence. 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 11 (Detailed balance condition).

Let $(X^{(t)})$ be the chain produced by the Metropolis-Hastings algorithm (see table 5.1). For every conditional distribution q whose support includes the support of ξ holds:

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 (5.4) on the kernel in equation (5.5).

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

$\alpha(x,y) = \min\left\{\frac{\xi(y)}{\xi(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}$, 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

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

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 fulfills the ergodic theorem (Theorem 10). To present this result in a more formal way the following convergence theorem is formulated.

Theorem 12 (**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*

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

- *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, $\mathcal{K}^n(x,.)$.

5.2.2 Gibbs Samplers

The simple but excellent principle of the Gibbs sampler is to use the true conditional distributions associated with the target distribution to generate samples from that distribution.

It is necessary that we can simulate from the conditional distribution $\xi_i(x_i|x_1,x_2,\ldots,x_p)$ $i = 1,2,\ldots,p$. 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 5.2: p-stage Gibbs algorithm

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

Example 2. *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 multi-stage 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.

5.2.3 Introduction to Reversible Jump MCMC

The method of reversible jump Markov Chain Monte Carlo (RJMCMC) has been introduced by Peter Green (see [25], [35]). It is in principle a generalisation of the Metropolis-Hastings method, which allows for jumps between spaces Θ_k of different dimensionality by defining a bijection (which is even a diffeomorphism) between well-constructed spaces which contain 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 fulfills

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

The appropriate kernel can be written as

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

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 fulfill 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 fulfilled:

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

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

$$\alpha_m(x,y') = \min\left\{1, \frac{f_m(y',x)}{f_m(x,y')}\right\} = \min\left\{1, \frac{\pi(dy')q_m(y',dx)}{\pi(dx)q_m(x,dy')}\right\} \tag{5.6}$$

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 $\mathbb{R}^{n_1}$ and $p(\theta^{(k_2)}|k_2)$ on $\mathbb{R}^{n_2}$. 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 $\mathfrak{C}_1$ 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 5.3: Reversible Jump algorithm

5.2.4 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 16 (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 [13] 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

$$\xi(x_I|x_{I^C}) \propto \xi(x) \tag{5.7}$$

$$\frac{\xi(x_I^{'}|x_{I^C}^{'})}{\xi(x_I|x_{I^C})} = \frac{\xi(x^{'})}{\xi(x)} \qquad for\ x_{I^C}^{'} = x_{I^C} \tag{5.8}$$

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

$$q(y_I|x_I,x_{I^C}) = \xi(y_I|x_{I^C}) \tag{5.9}$$

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 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 [44])

Theorem 13 (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.

Only in the case of the multistage Gibbs sampler a further statement regarding convergence can be made. For his we need the ***positivity condition***:
Let $(Y_1,\ldots,Y_n) \sim g(y_1,\ldots,y_n)$, then positivity of the marginal distributions, $m(y_i) > 0 \;\; i = 1,\ldots,n$, implies positivity of the joint distribution $g(y_1,\ldots,y_n) > 0$.
A sufficient condition for positivity is the following: all conditional distributions shall be absolutely continuous w.r.t. a dominating measure and all conditional and unconditional distributions have the same support which is connected. A connected support is not required for the Gibbs sampler in general, but unconnected parts can make the chain reducible and thus prohibit true convergence. Equivalence of the condition of absolute continuity to the positivity condition can be proven ([37]), its advantage is that it is easier to manage than the positivity condition.

Theorem 14. *If the positivity condition is fulfilled for a Gibbs Sampler, then:*

1. *ξ is the stationary distribution of the Markov chain.*
2. *For every measure μ the chain is ergodic, i.e.*

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

5.2.5 Theoretical consideration of Convergence

The sampler described in section (5.3) is a hybrid sampler containing some Gibbs-steps, a Metropolis-Hastings update and a mixture kernel of Gibbs and reversible jump update. Its composed transition kernel would look like that

$$\begin{aligned} \mathcal{K}_{hybrid} &= (\tfrac{1}{K}\mathcal{K}_{RJ}^{(\nu),(\phi_{n,g})} + \tfrac{K-1}{K}\mathcal{K}_{MH}^{(\nu),(\phi_{n,g})}) \circ \mathcal{K}_G^{(\tau)} \\ &\circ(\mathcal{K}_G^{(\lambda)} \circ (0.5 \cdot \mathcal{K}_G^{(I_g),(\beta_g)} + 0.5 \cdot \mathcal{K}_{RJ}^{(I_g),(\beta_g)})) \circ \mathcal{K}_G^{(p)} \end{aligned} \tag{5.10}$$

A short look shall be taken at the individual kernels of the algorithm and how much can be theoretically deduced according to the theory of section (5.1) for each kernel and their compositions.

At first we take a look at the Metropolis-Hastings type kernels, this shall also include the Gibbs steps, since each of them can be viewed as special type Metropolis-Hastings step. According to theorem 11, a sufficient condition for fulfilling the *detailed balance condition* and for the posterior to be the stationary distribution of the chain is that the support of the proposal distribution shall include the support of the posterior. This is naturally the case for the Gibbs steps, since distributions of the same family are involved. In the Metropolis-Hastings step for ν and $\varphi_{n,g}$ the support of the proposal distribution is per definitionem the set $\mathbb{R}^+ \times \mathfrak{N}$, the same as the support of the desired posterior. The reversible jump steps shall not bother us since these steps fulfill the detailed balance condition by construction. For each kernel, fulfilling the detailed balance condition is equivalent to the existence of an invariant density.

Another property that is required is *irreducibility*, which is fulfilled if the proposal distributions are positive on the support of ξ. This is again naturally the case since all proposal distributions are either continuous probability distributions with the same support as the posterior or in the case of ν positive by definition. As for statements about the composition of kernels, it is clear that each of the mixture kernels for β_g, I_g and $\nu, \varphi_{n,g}$ will inherit irreducibility from one of its components which has this property as stated above. 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 it also is the condition required for the Convergence theorem (12) to apply.

Regarding the cycle of kernels, irreducibility and aperiodicity is 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 just 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 master's thesis.

5.3 Application to the Student's t distribution model

Due to the choice of conjugate distributions many parameters can be updated by drawing from closed form distributions, which is the usual way for Gibbs sampling algorithm modules. In 2 cases a different approach will be chosen. Firstly the degrees of freedom parameter ν will be updated by a Metropolis-Hastings updating step, which in the special case of ν_{max} results even in a reversible jump step (see below 5.3.2). Secondly the coefficients β of the linear regression respectively ANOVA model and the indicator I_g are updated by Gibbs steps alternating with a reversible jump steps.
For the calculation of the full conditional distributions the common distribution of all modelled stochastic variables has to be derived:

$$p((I_g)_{g=1,\dots,G}, p, (\beta_g)_{g=1,\dots,G}, \nu, (\varphi_{n,g})_{n1,\dots,N;g=1,\dots,G}, \tau_\varepsilon) \propto \tag{5.11}$$

$$\begin{aligned} \propto\ & p(p)p(\tau_\varepsilon|c,d)p(\nu|K) \\ & \textstyle\prod_g p(I_g|p)p(\beta_g|I_g,\mu_g,\tau_g) \\ & \textstyle\prod_n p(\varphi_{n,g}|\nu)p(y_{n,g} - x_{n,g}^T\beta_g|\varphi_{n,g},I_g,\tau_\varepsilon) \end{aligned} \tag{5.12}$$

We will take the kernel in equation (5.10) apart and present the structure of the updates represented by the individual kernels in more detail. In order to have a better overview over the updates they will be grouped according to the underlying sampling scheme.

5.3.1 Gibbs Updates

Gibbs updates are used for three of the model parameters, the probability of differential expression p, the precision of each group β_g, λ, and the overall model precision, τ_ε. For a within-dimension update of the parameter β_g a Gibbs step is used as well. Their full conditional distributions can be calculated explicitly since we are working in a conjugate prior setting.

Update of τ_ε

The error of the model follows a Gamma distribution. The update will be drawn from:

$$\tau_\varepsilon|\dots \quad \sim \quad Ga(g + \frac{NG}{2}, h + \frac{1}{2}\sum_{n,g}\varphi_{n,g}(y_{n,g} - x_{n,g}^T\beta_g)^2)$$

Update of p

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

$$p|I \quad \sim \quad Be(a + i_1, b + (G - i_1))$$

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.

Update of λ

The hyperparameter λ determines the within-group precision for the β_g. Its updated value will be drawn from a Gamma distribution in the following way:

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

Update of β_g, within dimension

The following update is used for the vector of parameters which represent expression levels in the case of differential expression ($I_g = 1$) respectively the value that represents the overall expression level if there is no differential expression ($I_g = 0$):

(WD) update β_g conditional on all other variables

case $I_g = 0$

$$\begin{aligned}\beta_{g,0}|\lambda, \ldots &\sim N_1(\mu^*, (\lambda^*)^{-1})\\ & \mu^* = \frac{\tau_\varepsilon \sum_{n=1}^N \varphi_{n,g} y_{n,g} + \lambda\mu}{\lambda^*}\\ & \lambda^* = (\tau_\varepsilon \sum_{n=1}^N \varphi_{n,g} + \lambda)\end{aligned}$$

case $I_g = 1$

$$\begin{aligned}\beta_g|\lambda, \ldots &\sim N_S(\mu^*, (\Lambda^*)^{-1})\\ & \mu^* = (\Lambda^*)^{-1}(\lambda\mu + \tau_\varepsilon X D_{\varphi,g} Y_g^T)\\ & \Lambda^* = \lambda I_S + \tau_\varepsilon X D_{\varphi,g} X^T = diag(\lambda_1^*, \ldots, \lambda_S^*)\\ & with\ \lambda_s^* = (\tau_\varepsilon \sum_{i=1}^N \varphi_{n,g}^{(s)} + \lambda)\end{aligned}$$

According to equation (5.9) the Gibbs steps are straightforward to embed into the hybrid sampler, as we can sample from the full conditionals directly. For every single step the full conditional distribution is the transition kernel. This naturally implies properties which we require for the generic considerations regarding convergence of the algorithm. (see [37], 5.2.2 and 5.2.5 for more detail)

5.3.2 Metropolis-Hastings and Reversible Jump Updates

There are 2 situations where we require Reversible Jump Updates respectively a Metropolis-Hastings update.
Firstly a reversible jump move is required when the model for a gene changes between differential and non-differential expression. The parameter I_g determines the direction of the move, the parameter β_g has to be updated from the distribution in the new parameter space, as given by the reversible jump scheme (see table 5.3).
Secondly the update of the degrees of freedom parameter ν is performed by a Metropolis-Hastings step which simultaneously updates all the scaling parameters $\phi_{n,g}$. A reversible jump step is performed when the maximum value ν_{max} is reached and the model changes from a student's t error distribution with auxiliary variables $\phi_{n,g}$ to a Gaussian distribution.
Both of these updates represent a special setting of parameters where one parameter is conditional on the other and operates in a conjugate prior setting. A practical statement for the easier calculation of acceptance probabilities on a two (or more) components setting can be made when dealing with distributions which fulfill certain properties. A sufficient and also commonly used property is conjugacy of priors and a proposal distribution which equals the respective posterior distribution for the parameter θ_1 which is conditionally dependent on the other one θ_2. In this case the acceptance probability will only depend on the parameter θ_2. The reason why this is treated in such detail is that this is the setting used for all Metropolis-Hastings and Reversible Jump updates in this work.

Lemma 15 (**Acceptance probability**).

For a given model with likelihood $f(x|\theta = (\theta_1, \theta_2))$ and conjugate prior distribution $\pi(\theta_1|\theta_2)\pi(\theta_2)$ the acceptance probability of moving from $\theta^{(o)}$ to $\theta^{(n)}$ if the proposal density is the parameter's posterior density of θ_1 then the acceptance probability equals a term depending only on parameter θ_2

$$\min(1, \underbrace{\frac{n_{pr;\theta_1^{(n)}}(\theta_2^{(n)})n_{lh}(\theta_2^{(n)})}{n_{post;\theta_1^{(n)}}(\theta_2^{(n)})} \frac{n_{post;\theta_1^{(o)}}(\theta_2^{(o)})}{n_{lh}(\theta_2^{(o)})n_{pr;\theta_1^{(o)}}(\theta_2^{(o)})}}_{=\frac{m(x|\theta_2^{(o)})}{m(x|\theta_2^{(n)})}} \chi(\theta_2^{(n)}, \theta_2^{(o)})) \tag{5.13}$$

where $n_{pr;\omega}, n_{post;\omega}, n_{lh}$ stand for the normalisation parameters of the prior and posterior of the parameter ω respectively the likelihood function. $\chi(\theta_2^{(n)}, \theta_2^{(o)}))$ accumulates all parts of the acceptance probability originally depending only on the old and new value of θ_2.

Proof. The proof is based on the simple idea that regrouping certain terms will lead to a simpler expression.

$$\frac{\pi(\theta_1^{(i)}|\theta_2^{(i)})f(x|\theta_1^{(i)},\theta_2^{(i)})}{p(\theta_1^{(i)}|\theta_2^{(i)},\ldots)} = m(x|\theta_2^{(i)}), \qquad i \in \{o,n\}$$

Since we are in a conjugate prior distribution setting, all terms containing θ_1^i have to cancel out and the expression reduces to

$$\frac{n_{lh}(\theta_2^{(i)})n_{pr;\theta_1^{(i)}}(\theta_2^{(i)})}{n_{post;\theta_1^{(i)}}(\theta_2^{(i)})}$$

Any constant factors (independent of $\theta_2^{(i)}$) in the normalisation factor of the likelihood term will cancel out when calculating the fraction of the normalisation constants of the marginal distribution. All factors related to the update of $\theta_2^{(i)}$ only are merged in the factor $\chi(\theta_2^{(o)},\theta_2^{(n)})$. □

As in case of both our reversible jump updates only one type of move is possible for each of the parameter settings, the kernels of the reversible jump steps have the same structure as the Metropolis Hastings kernel. As they fulfill the detailed balance condition by construction, the existence of an invariant measure is implied. For more detailed considerations about the theoretical properties see [37], 5.2.5.

Update β_g, I_g, reversible jump

This step proposes a move between the two parameter spaces connected with differential expression respectively non-differential expression. There is a link between these updates and the Bayes test for differential expression. The structure of the variable A includes the Bayes factor for a test of 'differential expression' against 'no differential expression' of a gene, the proposals in the respective room encode the losses (inverse utilities) for each of these tests.

(RJ) case $I_g = 0 \rightarrow I_g = 1$: proposal for β_g

$$\begin{aligned} \beta_g|\varphi,\ldots \quad \sim \quad & N_S(\mu^*,(\Lambda^*)^{-1}) \\ & \mu^* = (\Lambda^*)^{-1}(\lambda\mu + \tau_\varepsilon X D_{\varphi,g} Y_g^T) \\ & L^* = diag(\lambda_1^*,\ldots,\lambda_S^*); \lambda_s^* = (\tau_\varepsilon \sum_{i=1}^{N} \varphi_{n,g}^{(s)} + \lambda) \end{aligned}$$

The auxiliary variable

$$\begin{aligned} A \quad = \quad & \frac{\prod_n p(y_{n,g} - x_{n,g}^T \beta_g | I_g = 1,\ldots)}{\prod_n p(y_{n,g} - \beta_{g,0} | I_g = 0,\ldots)} \\ & \frac{p(\beta_g|\mu_g, T_g, I_g = 1) p(I_g = 1)}{p(\beta_{g,0}|\mu_{g,0}, \tau_{g,0}, I_g = 0) p(I_g = 0)} \\ & \frac{p(\beta_{g,0}|I_g = 0,\ldots) p(I_g = 0)}{p(\beta_g|I_g = 1,\ldots) p(I_g = 1)} \\ = \quad & \lambda^{\frac{S-1}{2}} \sqrt{\frac{\lambda^*}{\prod_s \lambda_s^*}} \frac{p}{1-p} \\ & \mathbf{e}^{-\frac{1}{2}\varphi[(S-1) * \lambda\mu^2 - (\mu^*)^T \Lambda^* \mu^* + \lambda^* (\mu^*)^2]} \end{aligned}$$

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

case $I_g = 0 \rightarrow I_g = 1$: proposal for $\beta_{g,0}$

$$\begin{aligned} \beta_{g,0}|\varphi,\ldots \quad \sim \quad & N_1(\mu^*,(\varphi\lambda^*)^{-1}) \\ & \mu^* = \frac{\tau_\varepsilon \sum_{n=1}^{N} \varphi_{n,g} y_{n,g} + \lambda\mu}{\lambda^*} \\ & \lambda^* = (\tau_\varepsilon \sum_{n=1}^{N} \varphi_{n,g} + \lambda) \end{aligned}$$

The acceptance probability is α=min$\{1,A^{-1}\}$.

Update ν

As overdispersion, respectively underdispersion depending on the chosen value of the hyperparameter of the truncated Poisson distribution has caused trouble with the originally used Poisson prior, a uniform prior has been selected as an alternative. Additionally the uniform prior can adapt more flexibly to changes in the grid size of the underlying set. The algorithm allows the degrees of freedom to jump to the next higher or lower value within the ordered set $\mathfrak{N}$, instead of allowing jumps to any valid parameter value of the set; this is a simple Metropolis-Hastings step. Due to the advantage of a more straight forward implementation a rather easy updating algorithm has been chosen instead of investing too much time into the construction of a more sophisticated algorithm.

As an additional feature the commonly used Gaussian model (see [26], [23], [48]) was taken into account as well. As this is the standard distribution for any linear model we would like to keep it, unless the data requires a more robust model. For this purpose a reversible jump step will be introduced, which jumps between the t-model, consisting of a normal-gamma-model and the auxiliary variables $\varphi_{n,g}$, and a Gaussian model, which is equal to the upper model, when the $\varphi_{n,g}$ all equal one.

$$\begin{aligned}
A &= \frac{\prod_{n,g} p(y_{n,g} - x_{n,g}^T \beta_g | \varphi_{n,g}^{(n)}, I_g, \tau_\varepsilon)}{\prod_{n,g} p(y_{n,g} - x_{n,g}^T \beta_g | \varphi_{n,g}^{(o)}, I_g, \tau_\varepsilon)} \\
&\quad \frac{\prod_{n,g} p(\varphi_{n,g}^{(n)} | \nu^{(n)})}{\prod_{n,g} p(\varphi_{n,g}^{(o)} | \nu^{(o)})} \frac{p(\nu^{(o)} | \nu^{(n)}) \prod_{n,g} p(\varphi_{n,g}^{(o)} | \nu^{(o)}, \ldots)}{p(\nu^{(n)} | \nu^{(o)}) \prod_{n,g} p(\varphi_{n,g}^{(n)} | \nu^{(n)}, \ldots)} \\
&= \prod_{n,g} \frac{m^{(o)}(\nu^{(o)})}{m^{(n)}(\nu^{(n)})} \cdot \frac{p(\nu^{(o)} | \nu^{(n)})}{p(\nu^{(n)} | \nu^{(o)})}
\end{aligned}$$

where the second line results because of the conjugate prior setting for $\varphi_{n,g}$. the probability of selecting the new value $\nu^{(n)}$ given the old value $\nu^{(o)}$ is

$$p(\nu^{(n)} | \nu^{(o)}) = \begin{cases} 1 & \nu^{(o)} = 1 \vee \nu^{(o)} = \nu_{max} \\ 0.5 & else \end{cases}$$

thus resulting in the following expression (g^*, h^* see 5.3.2)

$$A = \left(\frac{p(\nu^{(o)} | \nu^{(n)})}{p(\nu^{(n)} | \nu^{(o)})} \right) \cdot \left(\frac{\frac{\nu^{(n)}}{2}^{\frac{\nu^{(n)}}{2}}}{\frac{\nu^{(o)}}{2}^{\frac{\nu^{(o)}}{2}}} \right)^{NG} \cdot \left(\frac{\Gamma(\frac{\nu^{(o)}}{2}) \Gamma(g^{*(n)})}{\Gamma(\frac{\nu^{(n)}}{2}) \Gamma(g^{*(o)})} \right)^{NG} \cdot \prod_{n,g} \frac{(h_{n,g}^{*}{}^{(o)})^{g^{*(o)}}}{(h_{n,g}^{*}{}^{(n)})^{g^{*(n)}}}$$

In the special case of the reversible jump step from t distribution to normal distribution the similar formula applies, the determinant of the additionally appearing Jacobian equals 1.

- For $\nu_{max} - c_{grid} \rightarrow \nu_{max}$:

$$A = \prod_{n,g} ({h^*_{n,g}}^{(o)})^{{g^*}^{(o)}} \frac{\Gamma(\frac{\nu^{(o)}}{2})}{(\frac{\nu^{(o)}}{2}^{\frac{\nu^{(o)}}{2}})^{NG}\Gamma({g^*}^{(o)})}$$

- For $\nu_{max} \rightarrow \nu_{max} - c_{grid}$:

$$\prod_{n,g} ({h^*_{n,g}}^{(n)})^{-{g^*}^{(n)}} \frac{\Gamma({g^*}^{(n)})(\frac{\nu^{(n)}}{2}^{\frac{\nu^{(n)}}{2}})^{NG}}{\Gamma(\frac{\nu^{(n)}}{2})}$$

Update $\varphi_{n,g}$

The auxiliary variable $\varphi_{n,g}$ is drawn from the following Gamma distribution:

$$\begin{aligned} \varphi_{n,g}|\ldots &\sim Ga(g^*, h^*) \\ g^* &= \frac{\nu+1}{2} \\ h^*_{n,g} &= \frac{1}{2}(\nu + \tau_\varepsilon(y_{n,g} - x^T_{n,g}\beta_g)^2) \end{aligned}$$

5.4 General MCMC Convergence Analysis

Before presenting results of such convergence diagnostics, we provide an overview over the involved notions and statistics (for more details see [19]). Since all MCMC algorithms have to be terminated after a certain number of samples draws, determining whether it is safe to assume that the obtained samples are truly representative of the underlying distribution is important. As the samples are drawn from Markov chains they are generally correlated. This correlation makes the draws less informative than iid draws from the stationary distribution would be. Thus the process of exploring the stationary distribution will be slowed down.

The theoretically more well-founded attempt of analysing the transition kernel itself in order to determine the number of iterations required has proven not to be very fruitful for practical use, as the resulting bounds were quite loose and too large to be of practical value. Thus the more common approach is applying various forms of diagnostics tools to the algorithm's output in order to come to conclusions a posteriori. Based on this prior information the algorithm is run for the given number of draws which will likely ensure convergence. A critical issue for all these diagnostics is that however one might construct a statistic, one cannot compare sample distributions to the unknown stationary distribution, but only to other sample distributions (either from different iterations or from different parts of the same chain). Thus 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 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 estimate of quantiles of functions of the analysed parameters. For the calculation of the diagnostics with a given precision a minimum number of draws, N_{min}, is required under the assumption that they were independent.
  The aim of the method is the estimation of a quantile q with accuracy r, which has to be attained with probability s, $\mathbb{P}[q - r \leq \hat{\theta} \leq q + r] \geq s$.
  The output will be the total number of iterations to be run in order to fulfill 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, as it describes how many samples have to be discarded in order to consider the k-th one an iid draw from the stationary distribution.

- ***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 fulfilled, 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 fulfilled 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***
  Autocorrelation describes the correlation between different time points of a time series.
  Definition 17 (Autocorrelation function).

  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. The behaviour of time series, particularly ones generated from MCMC samplers, can be described by an *autoregressive process.*
  Definition 18 (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. 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.

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 sample from a Markov chain
Partial autocorrelation	graphical	Correlation of sample from a Markov chain

Table 5.4: Overview over diagnostics and some properties; all these methods have in common that they are generally applicable to any MCMC algorithm, only take single chains into account and work for univariate parameters (see [19] for more details)

Several of these diagnostics have been implemented in the R package coda 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 nullhypothesis 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 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 lengthes 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.

6 Computational Results

Before applying the algorithm to biological data where nothing is known about the underlying distribution we wish to explore, we have to test the program with artificial data sets in order to perform a primary sanity check. Besides this primary goal, two other aspects are of importance. One is the definition of state space for ν, especially the grid length and its influence on the outcome as well as convergence results. Another is convergence analysis the possibility of getting an estimate of how many draws will be necessary to get sufficiently large samples from the invariant distribution.

6.1 Test Data Sets

a)

Set	μ_1	μ_2	ν	%
1	-5	5	∞	50
1	-1	1	∞	50
2	-5	5	4	50
2	-1	1	4	50
3	-5	5	10	50
3	-1	1	10	50
4	-5	5	∞	20
4	-1	1	∞	80
5	-5	5	4	20
5	-1	1	4	80
6	-5	5	10	20
6	-1	1	10	80

$\nu = \infty$ represents the normal distribution.

b)

i	$\mu_{i,1}$	$\mu_{i,2}$	%	
1	-5	5	10	
2	-1	1	40	
3	-0.1	0.1	10	
4	-0.01	0.01	40	
1	-2.5	2.5	10	10
2	-1	1	10	20
3	-0.5	0.5	20	
4	-0.01	0.01	10	30
5	-0.001	0.001	10	20
1	-5	5	10	
2	-2.5	2.5	10	10
3	-1	1	10	20
4	-0.1	0.1	10	10
5	-0.01	0.01	10	40
6	-0.001	0.001	10	10

$\nu \in \{4, 10, \infty\}$

Table 6.1: Test data sets

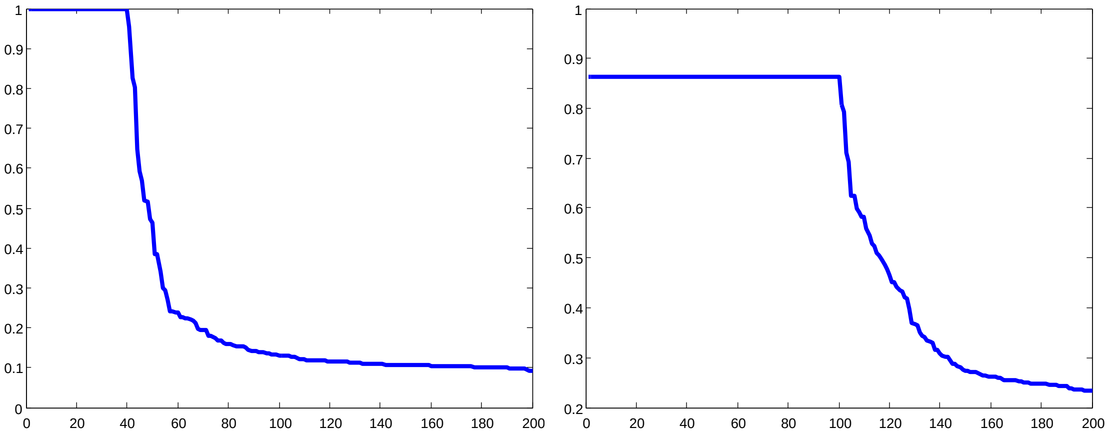
For testing purposes 2 types of data sets have been generated. The first was designed as a sanity check for a generic case where 50 % of genes are differentially expressed and the other 50 % are not, as well as a case of 20% vs. 80% differential vs. nondifferential expression setting (see 6.1 a)). 100 respectively 200 artificial genes have been sampled

from the respective distribution. The aim of the test was to see whether t-distributed and normal data could be told apart, and if the t-parameter of the underlying and in this case known distribution could be estimated accurately.
The second data set creates a more diverse range of genes with 4 to 6 stages ranging from 'not differentially expressed' to 'clearly differentially expressed'. The table (6.1 b)) show the values for the group means that have been tested. The generated data follows $N(\mu_{i,j}, 1)$ and $t_\nu(\mu_{i,j}, 1)$ $(i = 1, \ldots, 4/5/6, j \in \{1, 2\})$ respectively.

6.2 Results

6.2.1 Data Set 6.1 a)

For the first data set and all types of distributions the set $\mathfrak{N}^{(1)}$ with $c_{grid} = 1$ and $\nu_{max} = 45$ has been used as parameter space for ν. It is not necessary to choose a smaller step size, as this first check should only provide an indication, if the algorithm chooses the right direction with respect the degrees of freedom. Once the algorithm is close enough to its final error model, the acceptance probabilities become very small and keep the chain from moving around. This is problematic for the mixing behaviour, but since accurate estimation of the posterior is not our goal in this test, it is not a serious problem. However this example means to show that keeping to a relatively large grid size for the whole simulation can be problematic and thus will be avoided in the following test runs.

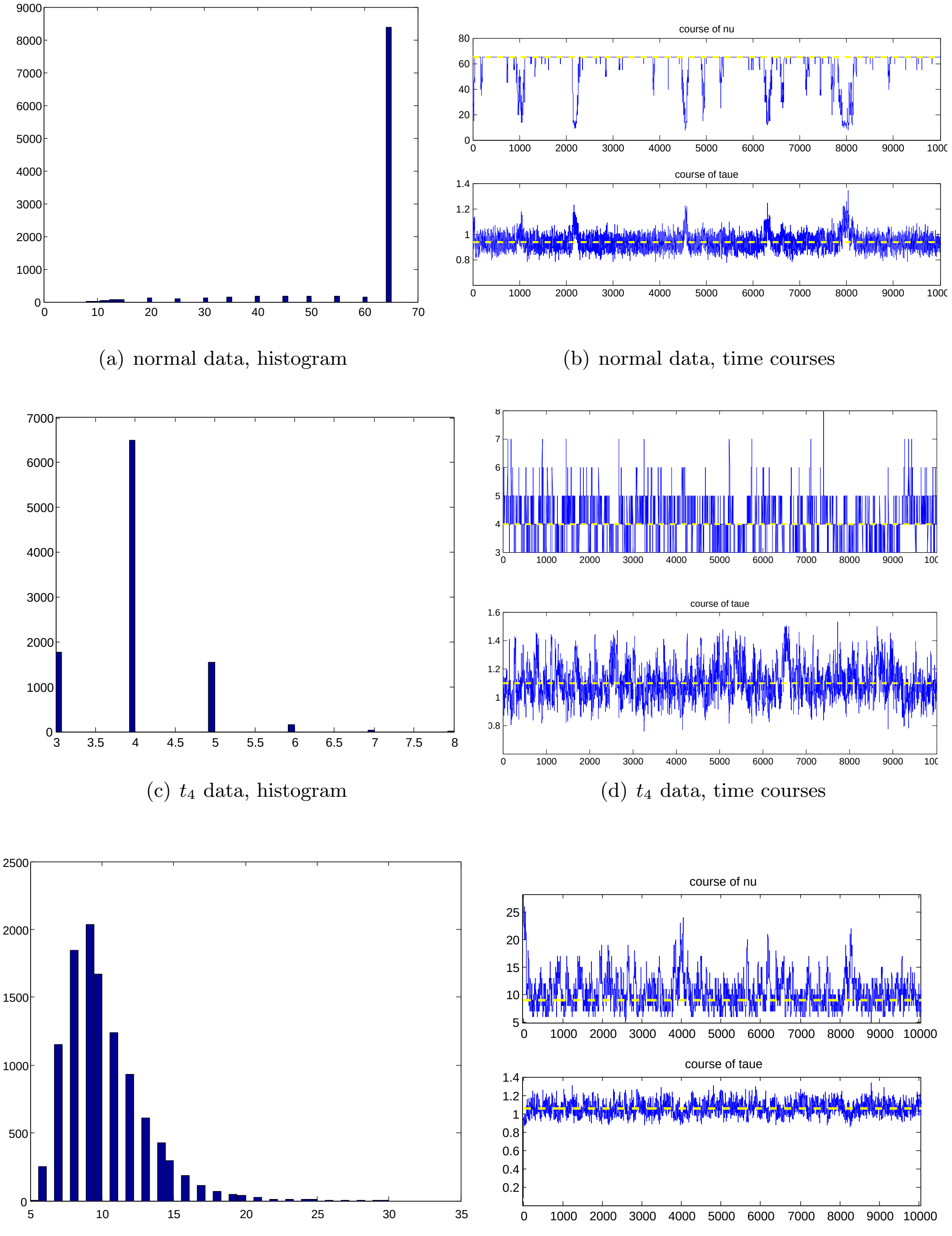


(a) 20% differentially expressed data

(b) 50% differentially expressed data

Figure 6.1: Ranked gene function plots for normal data according to table 6.1,a)

As one can see in the figure 6.1 all simulated 'highly differentially expressed' genes are indeed recognised as 'highly differentially expressed'. This is a first sanity check that the program has the ability to identify such gene. However a trend towards staying in



(a) normal data, histogram

(b) normal data, time courses

(c) t_4 data, histogram

(d) t_4 data, time courses

(e) t_{10} data, histogram

(f) t_{10} data, time courses

Figure 6.2: Histogram of ν and time courses for ν and τ_ε for the data from 6.1,a). The grey dashed line marks the mean of τ_ε, respectively the median of ν.

the preferred model space in clear cases and hardly ever moving by chance is visible, as the acceptance probability for such a move is extremely low. This mixing problem can be arguable for the unambiguous ranking of genes, as several genes might have p-values equal to 1 or 0. Graphics for the 2 types of t-distribution look similar and are therefore not included.

The histograms in Figure 6.2 clearly show that the algorithm is able to discern the underlying distribution even when the wide grid of $c_{grid} = 1$ is used. The time course of the degrees of freedom parameter for the t_{10} data shows why the choice of 45 as cut-off value for normality was sensible. In order to stay within a t-distribution setting and clearly discern it from a normal distribution setting the choice of a value of ν_{max} above 35 was required. This stays true for t distributions with 20 or more degrees of freedom, however 10 degrees of freedom can be viewed as the maximum value for providing results in ranking significantly different from the normal distribution setting. However an improved resolution of different degrees of freedom can only be obtained with a finer grid over ν. This is discussed in the next section.

6.2.2 Data Set 6.1 b)

For the data sets constructed according to table 6.1 b), two sets will define the state space of ν:

- $\mathfrak{N}^{(BI)}$, only used during a first burn-in phase has a grid length $c_{grid} = 1$ to roughly determine the direction ν will take. The test runs on the data from 6.1 a) have shown already that the algorithm does this accurately.
- $\mathfrak{N}^{(SA)}$ is defined by a finer grid length of $c_{grid} = 0.01$ and will be used for the second burn-in phase on the fine grid, as well as updates on this grid. The reason for using a refined grid is the possibility to better explore the posterior distribution. For such a small grid size the resulting posterior can be viewed as discretisation of a smooth density, which is important for the mixing of the chains. It also allows the algorithm to perform jumps in cases when the acceptance probability would get too small for integer steps.

Figure 6.3 illustrates the effect of the underlying model on the ranking of genes. It shows results of different scenarios for the distributions for 200 genes. Normally, t_4 and t_{10} distributed data are fitted by their underlying distribution, as it was found by the algorithm. Additionally for the t-distributed data sets a normal model has been fitted in order to visualise the influence the model has on the ranking.
A trend is visible for these models. The normal distribution model on Gaussian data behaves differently than t distribution and normal distribution on t distributed data, as in any case the ranking curve for Gaussian date intersects the curves for t data. The overall behaviour meets our expectations that a t distribution with higher degrees of freedom has a curve closer to the normal distribution than the t distribution with lower degrees of freedom. The more genes are used for the runs, the less apparent the difference becomes. Additionally it can be stated that an incorrectly fitted Gaussian model differs from the student's t distribution, independently of how many genes are provided as input

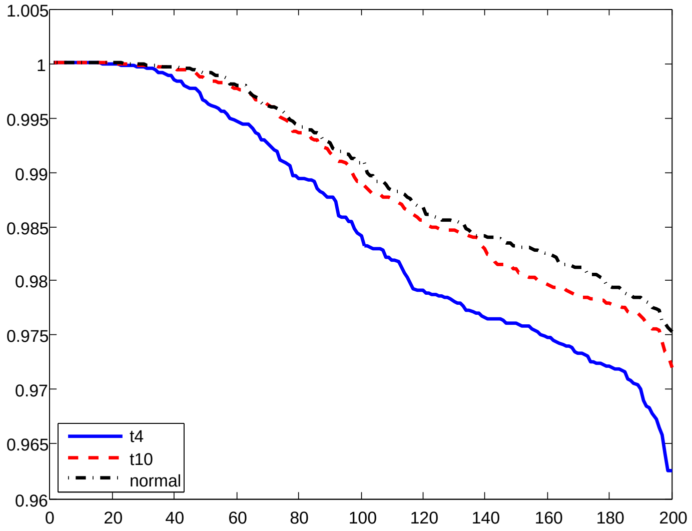


Figure 6.3: Ranked gene function plots for normal, t_4, t_{10} data for set 2 according to table 6.1,b)

data. However the significance of that difference is dependent on the respective data and cannot be generalised.

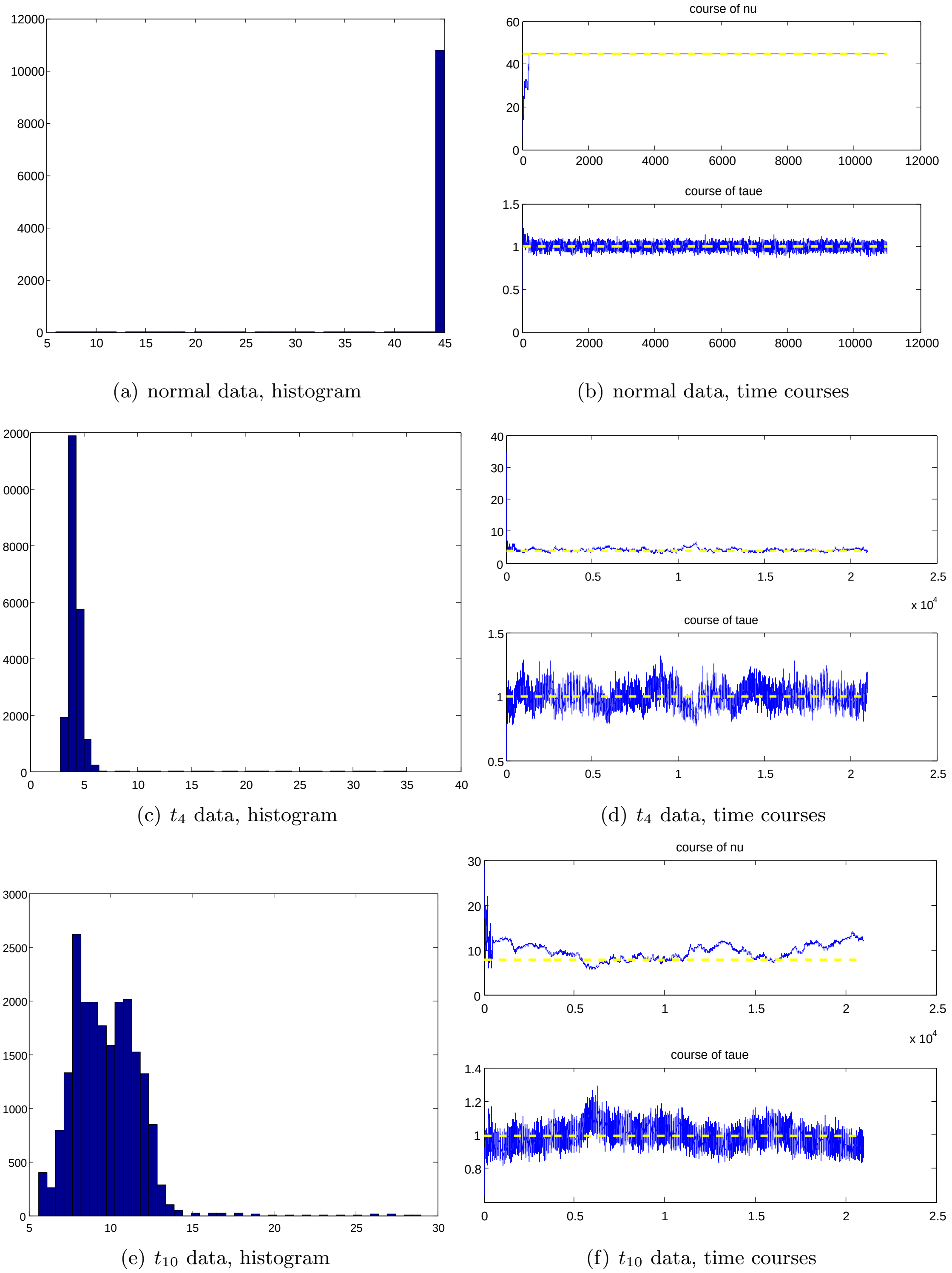
For normal data, t_{10} and t_4 data the behaviour of ν is as described above: ν tends more or less directly towards its true value, or a value close by. For the normal data set the value very rarely leaves the normal distribution setting, as the acceptance probability for the jumping to the higher dimensional t model is extremely low. On the t_4 data set the algorithm shows the best behaviour in the sense of model selection, as the variation of ν around its median value is very low. In contrast to this, the t_{10} data set has high variation around the median of about 10. Additionally the draws of ν point towards a bimodal posterior distribution.
Due to the finer grid the degrees of freedom stay in the range of ν which is approached during the first burn-in phase. The following sampling steps explore the posterior distribution around this value more precisely.

6.3 Convergence Assessments for the data

A practical analysis of the resulting Markov chains has been conducted using the coda package for statistics program R ([2]). The package provides a large inventory of diagnostics tools focusing on various aspects of convergence assessment.

Several aspects have been regarded in this analysis:

- Stationarity
- autocorrelation



(a) normal data, histogram

(b) normal data, time courses

(c) t_4 data, histogram

(d) t_4 data, time courses

(e) t_{10} data, histogram

(f) t_{10} data, time courses

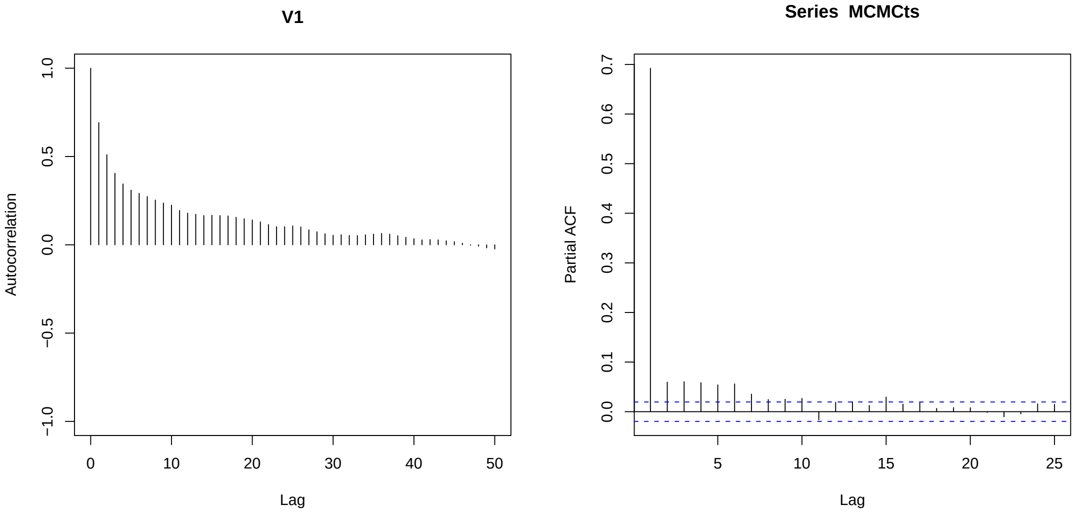
Figure 6.4: Histogram of ν and time courses for ν and τ_ε for the data from 6.1,b) 2. The grey dashed line marks the mean of τ_ε, respectively the median of ν.

- estimates of Burn-In length and required number of iterations

The analysis will be presented for the data of set 6.1 a, with 20% - 80% ratio of highly differentially expressed genes.
Since the variable ν is the parameter of interest, convergence analysis has been conducted focussing on ν. The analyses will be conducted for the t_4 and t_{10} data which have been selected for two reasons. Firstly these are the cases that also occur in the biological data sets in the next chapter and thus allow us direct comparison and conclusions. Secondly the analysis of the normally distributed data yields no further gain of knowledge and is directly comparable to the t-distributions.

	t_4 data	t_{10} data
Effective size	668	272
Required iterations	1398	7176

Table 6.2: Effective size of the 10000 samples and required number of iterations according to Raftery-Lewis ($q = 0.5, r = \pm 0.05, s = 0.95$)

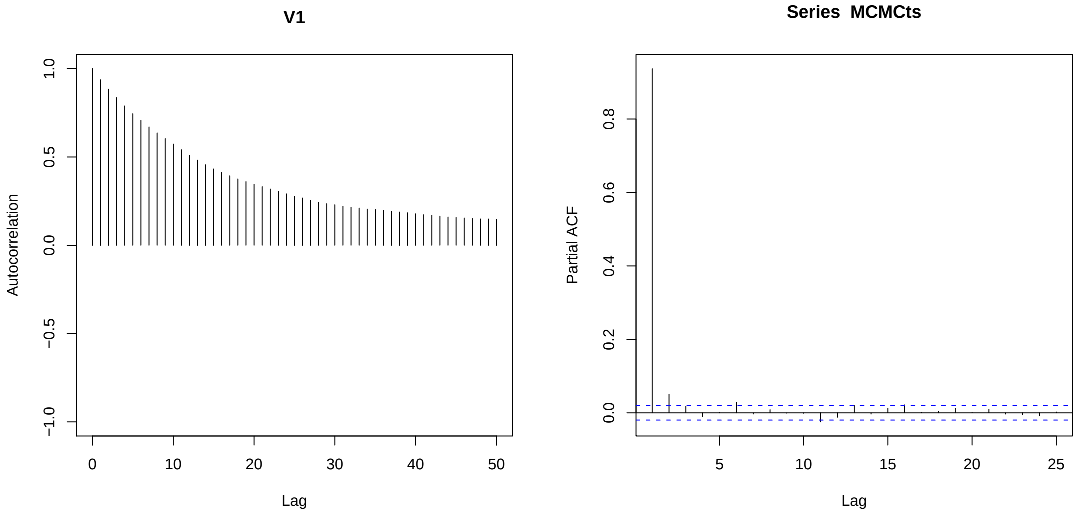


(a) Plot of the Autocorrelation function

(b) Plot of the Partial Autocorrelation function

Figure 6.5: Graphical convergence diagnostics for t_4 data

The Heidelberger-Welch test compares the second half of the data to the first 50%, a procedure also referred to as half width test, if this should fail, it discards the first 10%, 20%, etc. and continues the comparison until either the null hypothesis is accepted or the limit of 50% is reached, which automatically implies absolute failure or in other words that the chain has not converged yet. The chains for all 3 types of distributions have passed the test based on the Heidelberger-Welch diagnostic instantly without requiring



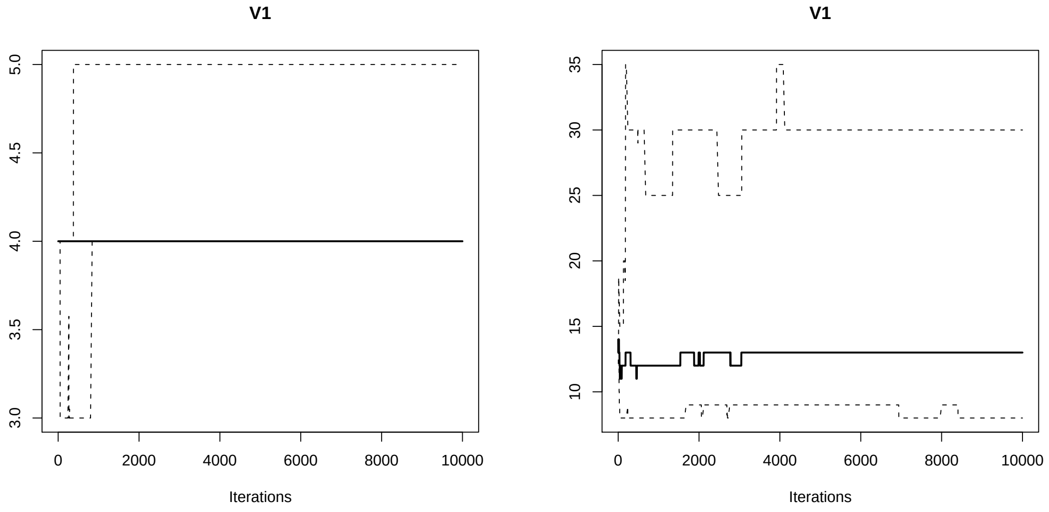
(a) Cumulative estimates of quantiles (0.025, 0.5, 0.975)

(b) Plot of the Autocorrelation function

Figure 6.6: Graphical convergence diagnostics for t_{10} data

the elimination of values.
The estimated effective size of the chains is only a small fraction of the original sample size, as the degrees of freedom ν are just integer-valued in this case. This causes huge autocorrelation within the chain of the degrees of freedom values of the t-distribution and gives us another reason, why such a coarse grid should be avoided.

As a result of the autocorrelation slow mixing can be seen as a problem for estimating a density in case of having a coarse grids. However it is not a serious problem, as convergence occurs very rapidly during the phase of burn-in towards the value where the main probability mass of the density will likely be located. For recognising an underlying t distribution it is sufficient to know that the degrees of freedom lie within a few integers around 4 respectively 10-15 with 95% probability. The graphical representation of the first values of the autocorrelation function shows that large values are taken, due to the underlying jumps between integers. This is another argument for choosing a finer grid than $c_{grid} = 1$.
In the plot of the cumulative estimates of quantiles (Figure 6.7) no clear trend is visible, which would give an indication that the stationary distribution has not been reached yet. For the t_4 distribution there is no visible change after the first 1000 draws. What can also be deduced from the graph is that estimation of any other quantile than the median of ν is not reasonable in a case like this, as even the 2.5% quantile collapses with the mean, as only integer values are available. The t_{10} results show no real trend either, fluctuations in the outer quantiles occur in the beginning, but cease to have influence once that enough samples have been drawn. It shall not go unmentioned that even in case of the t_{10} data the true value of ν of the underlying distribution lies within the empirically estimated



(a) t_4

(b) t_{10}

Figure 6.7: Cumulative estimates of quantiles (0.025, 0.5, 0.975) for t_4 and t_{10} data

95% interval. In the case of a finer grid things will look quite differently as the posterior distribution can be regarded as continuous then. Thus quantile estimates make much more sense for getting an impression of actual convergence.

To summarise the results of the last few pages we can say that convergence does generally occur within the chosen number of draws. This is confirmed by the Heidelberger-Welch test which all chains have passed. The autocorrelation plots show quite drastically that slow mixing occurs for the chains of ν. The effective sample size supports this statement. For both data sets it is only a few hundred draws out of 10000 draws over all which can be viewed as iid and used for estimates of e.g. quantiles. This is caused by the coarse grid on the state space of ν which keeps the chain from moving more freely around the median estimate of ν. This is the reason why for the biological data sets a finer grid for ν will be used in order to avoid such difficulties and be able to get reliable estimates for the posterior distribution of ν.
The comparison of rankings for true underlying t distribution against incorrectly fitted normal distribution and against a true underlying normal distribution has revealed that differences between the true underlying distributions decreases with increasing number of genes analysed. The differences between the results of fitting correct model distribution and the incorrectly fitted error model remain independent of the gene number. However their significance is dependent on the individual sample and cannot be generalised.

7 Biological Data Sets

We want to test the algorithm not just on artificial data sets, but also in 'real life' situations of microarray data analysis. Three data sets are selected which cover some of the variety possible for such experiments. At first an overview over the design of these data sets used as examples shall be given. Then the results of the Matlab runs are presented as well as the convergence assessment using the coda package.
Additionally to the methods of convergence assessment presented in the previous chapter a comparison of parallel chains will be conducted. We use four chains with widely dispersed starting points in order to get a better impression of whether one of the chains gets stuck at a local modus.
The simulation results for determining the validity of a student's t distribution assumption for the data are compared to runs which are based on a (fixed) normal distribution model. For a purely statistical comparison we take a look at resulting gene rankings which represent the a posteriori inference of the indicators I_g. In order to determine a possible difference in biological conclusions, the biological effects of both methods are compared by means of Gene Ontology terms.

At first the numerical setting of the tests is described. As already explained in the previous chapter, we work in a setting, where the state space of ν will be defined by the two sets:

- $\mathfrak{N}^{(BI)}$, only used during a first burn-in phase has a grid length $c_{grid} = 1$ in order to roughly determine the direction ν will take, while
- $\mathfrak{N}^{(SA)}$ is defined by a finer grid length of $c_{grid} = 0.05$ and will be used for the second burn-in phase on the fine grid, as well as updates on this grid. The reason for using a refined grid is the possibility to explore the posterior distribution in a better way. It also allows the algorithm to perform jumps in cases when the acceptance probability could get too small for integer steps.

The burn-in lengths have been chosen as 500 and 11000 draws have been performed in total. The initial values of ν have been randomly drawn from the set of $\{15, 20, \dots, 40\}$. All other variables have been initialised by randomly drawing from their prior distributions.

7.1 Data Sets

The chosen data sets exemplify some variety of microarray experiment settings. All details important for the following analysis will be presented very briefly.

Data Set	Fly 'Gender Comparison'	Mouse testis data	Endothelial cell data
experiment type	group comparison	time course	time course
Species	Drosophila melanogaster	Mus musculus	Homo sapiens
array type	cDNA	Affymetrix	CodeLink Human Uniset
number of arrays	6	22	24
Design	male female Cy3 (3×) Cy5(3×) Cy5 (3×) Cy3(3×)	11 time points 2 measurements per time point	8 time points 3 measurements per time point

Table 7.1: Overview over biological data sets

- ***Drosophila 'Gender Comparison'***

  This data set represents one of the classical experimental settings, i.e. the comparison of one type of individuals against another, in our case this will be male vs. female. The data set is taken from a set of experiments used for the paper [43]. On 6 cDNA arrays the expression of 13,826 genes is measured for both types of individuals, male and female fruit flies. The groups are coded by dye colours which makes swapping the dyes necessary in order to take care of effects introduced by measuring the respective light intensities, i.e. for half of these arrays the dye colours are switched.

- ***Affymetrix time course of Mouse testis cells***

  The data comes from a series of experiments conducted at the Griswold Lab by Shima et al. [41]. This time course means to track the gene expression in mouse testis cells during the progression of spermatogenesis. Samples are collected at 11 time points and for each time point 2 arrays are used for the evaluation in order to take care of some biological variation. Affymetrix chips are used for the experiments which have just one channel to be considered.

- ***Endothelial cell apoptosis time course***

  The data has been taken from experiments conducted by Affara et al. [4]. This data set holds another time course which observes the death of human endothelial cells. Sample collection takes place at 8 time points, from the time of induction of apoptosis in the cell to 24 hours after that. 3 measurements for each time point are included in the analysis, these come from independent repeats of the experiment. The RNA has been hybridised to CodeLink Human Uniset 20K gene arrays.

7.2 Matlab Results & Coda Analysis

For the analysis the expression values were extracted using the tool FSPMA (cite?). For normalisation the vsn transformation for variance stabilisation and normalisation of package vsn in R has been used.

Two main findings can be reported when looking at numerical results. Firstly for all data sets a student's t model with degrees of freedom of around 4 and about 10 is chosen. Secondly the inference results for this student's t model differ from those of a normal distribution model and have some significant impact on the biological outcome.

Before discussing the results in more detail a brief overview over the convergence assessment will be given. In order to draw any conclusions about the outcome, it should be clear that these come from the stationary distribution. All 4 chains of all 3 data sets have passed the Heidelberger-Welch halfwidth test. This result implies the equality of sample means of the first and second half of each chain. Furthermore the shrinking factor of the Gelman-Rubin statistic has been calculated which is almost 1 for all data sets. This statistic determines the homogeneity behaviour of the different chains, and values close to 1 mean that there is hardly any difference between the chains. Thus pooling the data for the calculation of estimates, e.g. the sample mean, for all chains is justified. An overview over the time course of the Gelman-Rubin diagnostics of all 3 data sets is presented in the Figure 7.1.

The effective size of the samples is quite small for all data sets. This is related to the high autocorrelation between the draws, as the degrees of freedom parameter is varying, once it is close to its preferred value. Although this is a nasty behaviour, if a good estimate of the posterior density were required, this is not an unwanted effect in terms of model selection, as the degrees of freedom parameter is determined very precisely by our model. Additionally this behaviour has already been observed for the test data sets of student's t distributions.

	Flies Data	Mouse data	Human data
Effective size	242	230	220
Required iterations	6414	8132	8240
Shrinking factor	1.05	1.06	1.00

Table 7.2: Shrinking factor of the combined chains, the average effective size of the 10000 samples and average required number of iterations according to Raftery-Lewis

Our first observation is that all data sets tend towards a student's t distribution model. This is little of a surprise since overdispersion is a well-known problem for microarray data. Conveniently, even though unintendedly the observed t distributions have degrees of freedom ν of about 4 and about 10, thus being directly comparable to the test data w. r. t. convergence properties. For both the drosophila data set and the human endothelial cell data set a distribution with degrees of freedom between 4 and 5 is found to be the best error model a posteriori, as can be seen from the graphs in figure 7.2. In the case of the mouse data set a trend towards a degrees of freedom value of circa 10 is obvious for all 4 chains. Thus in all three cases the data would require an error distribution which clearly has more probability mass in the tails than the normal distribution.

Secondly and more interestingly the behaviour of gene ranking is not only found to be statistically different when compared to ranking computed under the normal distribution assumption, it also is consistent with the observations of the test data sets regarding this aspect. For all three data sets the t distribution behaves more conservatively w.r.t. the majority of genes than the normal approximation. As we can see in the graph of the Fly data (figure 7.2) in contrast to the time course data sets a high number of genes is differentially expressed. For this high portion of differentially expressed genes the effect of the *t* distribution is visibly larger than for the other two data sets.

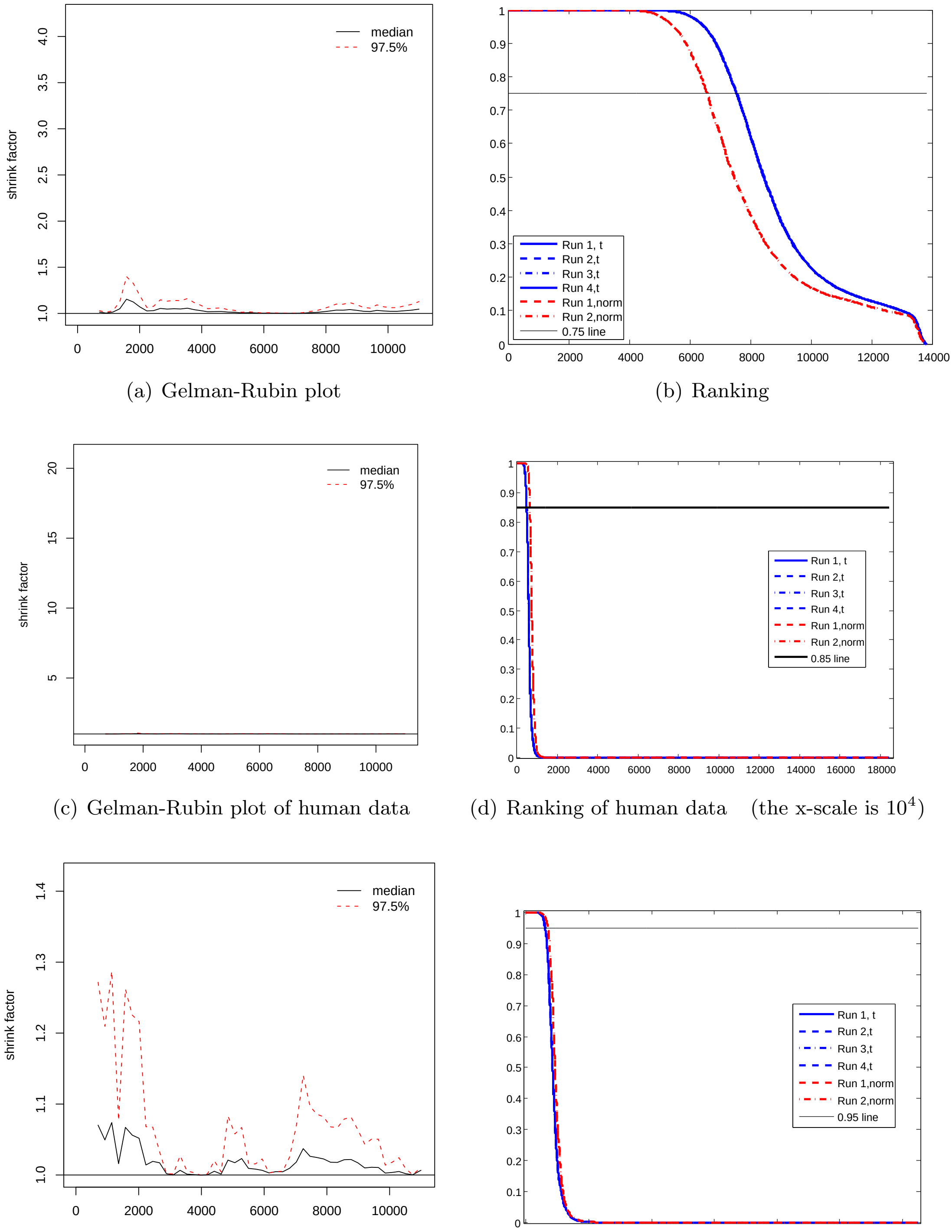
A less steep slope and thus higher number of recognised differentially expressed genes not only affects the amount of genes taken into account for further investigation, but also has an impact on biological conclusions. For this purpose Gene Ontology terms which are linked to the observed genes have been tested. This is a simple way of classifying the data w.r.t. certain biological aspects which have a more concrete interpretation than the genes or gene products themselves.

The Gene Ontology method is based on the idea of defining biological processes, molecular functions or cellular components within different levels of generality which are linked to certain genes. These terms are interconnected with specific terms of the next higher or lower level according to their functionality which allows for a graphical representation in form of a DAG.

For each Gene Ontology term a Fisher's exact test for the amount of genes related to a biological process within the 2 sets of genes is performed. In our case we test the set of differentially expressed genes against the rest of the genome. An adjustment of p-values is required for the multiple tests performed in this step. The webtools FatiGO [5] and DAVID uses the FDR (false discovery rate) method by Benjamini and Hochberg for this purpose, other correction methods are possible as well.

We compare the amount of GO terms which are significant for the inference result of the t distribution model with the amount found for the Gaussian model. An overall trend can be observed for all 3 data sets. The t distribution is more conservative than normal distribution in the following sense. If few genes are differentially expressed, less differentially expressed genes and less significant GO terms are found for the t distribution model. However these terms are usually the same terms as the ones with lowest p-value found for the normal distribution model results. If a high amount of genes is differentially expressed, the t distribution tends to find more GO terms.

In the fly data set a high amount of genes is highly differentially expressed between males and females and the student's t model tends towards higher posterior probability of differential expression. For such large amounts of differentially expressed genes, the usual GO analysis based on the hypergeometric distribution tends to underestimate the number of associated GO terms. For his reason the GO term inference was done with a counts based test implemented by Peter Sykacek in Matlab, which uses a Binomial distribution with probability 0.5 as null hypothesis that finding active and inactive genes



(a) Gelman-Rubin plot

(b) Ranking

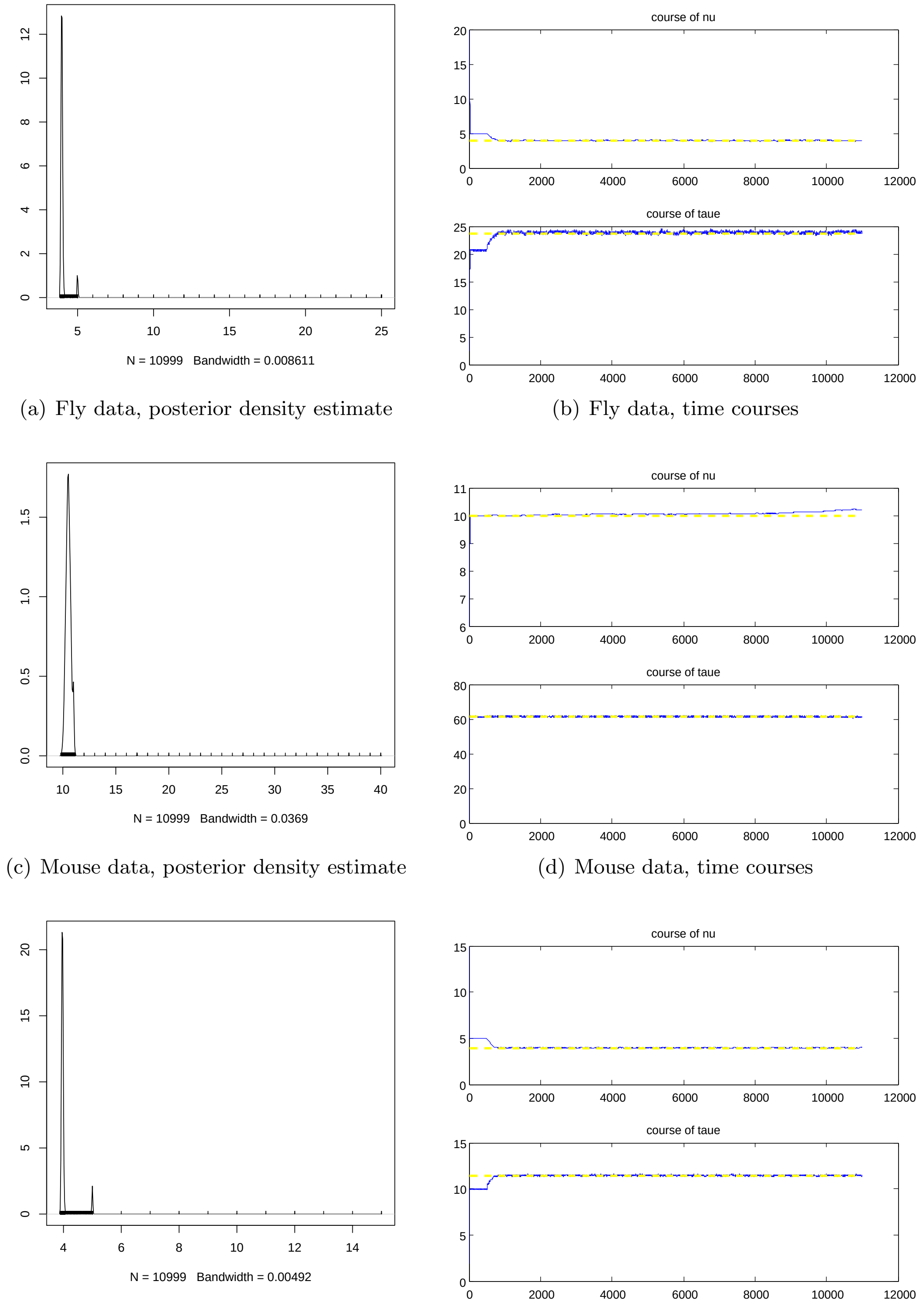
(c) Gelman-Rubin plot of human data

(d) Ranking of human data (the x-scale is 10^4)

(e) Gelman-Rubin plot for mouse data

(f) Ranking for mouse data

Figure 7.1: Gelman-Rubin plot and ranking of t distributed and normal model for mouse data



(a) Fly data, posterior density estimate

(b) Fly data, time courses

(c) Mouse data, posterior density estimate

(d) Mouse data, time courses

(e) Human data, posterior density estimate

(f) Human data, time courses

Figure 7.2: Histogram of ν and time courses for ν and τ_ε for the 3 data sets. The grey dashed line marks the mean of τ_ε, respectively the median of ν.

for inactive GO terms is equally probable and a onesided alternative that active genes are enriched.
With this method, 72 enriched GO terms of biological processes have been found for the student's t model and only 49 for the Gaussian model. Many processes related to cellular metabolism and biosynthesis (e.g. "cellular metabolic process" (GO:0044237), "RNA metabolic process" (GO:0016070)) are significantly enriched. Even clearer in terms of differing between male and female flies are processes like "reproduction" (GO:0000003), "reproductive process" (GO:0022414), "sexual reproduction" (GO:0019953), "gamete generation" (GO:0007276), "female gamete generation" (GO:0007292) and "oogenesis" (GO:0048477) which are all found significantly enriched for both the student's and the Gaussian model. A number of processes which are only found for the student's t model are related to alternative splicing, such as "RNA splicing" (GO:0008380), "RNA splicing, via transesterification reactions" (GO:0000375), "RNA splicing, via transesterification reactions with bulged adenosine as nucleophile" (GO:0000377) or "nuclear mRNA splicing, via spliceosome" (GO:0000398). Several experiments (see e.g. [33]) have described that alternative splicing is important for regulating sexual differentiation in fruit flies.

For the analysis of the mouse data set the webtool FatiGO [5] has been used. We have included all genes with a posterior probability of being differentially expressed greater then 0.95. 42 GO categories for biological processes have been identified as significant in the set of the highest ranked genes found by the Gaussian model, 24 such categories have been identified for the respective set of the student's t model. As we are observing the development of mouse testis cells, we would expect a high activity of Gene Ontology terms related to the development of gametes. This expectation is met by both data sets, as terms like "sexual reproduction" (GO:0019953), "gametogenesis" (GO:0007276), "fertilization" (GO:0009566), "male gamete generation"(GO:0048232), "spermatogenesis" (GO:0007283), "sperm-egg recognition" (GO:0035036) are found to be significant, these finding are similar to those described in the original paper by Shima et al. [41]. The student's t model very specifically finds almost only terms related to these in GO. In case of this data set the t distribution provides a very specific focus which might not be a disadvantage.Additionally the normal distribution model finds some biological processes which can be of interest, such as "tube development" (GO:0035295), "actin cytoskeleton organization and biogenesis" (GO:0030036) or "skeletal development" (GO:0001501).

We have used the webtool DAVID [22] for the analysis of the endothelial cell data. For the Gaussian model 84 GO terms of biological processes and 4 cellular components were significant, whereas for the student's t model 35 biological processes and 3 of cellular components could be found. Here the difference between the results are more significant than for the other data sets. For both models the "nucleus" (GO:0005634) and "nucleolus" (GO:0005730) are found to be important cellular components which is consistent with knowledge that nuclear condensation is an event during apoptosis. Among the significant biological processes terms like "negative regulation of cellular process" (GO:0048523), "negative regulation of biological process" (GO:0048519) can be found which are not unexpected, as downregulation of normal processes comes along with cell death. Also found for both models are processes linked to cellular development, such as "developmental process" (GO:0032502), "cellular developmental process" (GO:0048869). Additionally, only the Gaussian model finds several terms related to mitosis, such as

”mitosis” (GO:0007067) and ”mitotic cell cycle” (GO:0000278) which is also found by Affara et al. [4] and cell death, ”death” (GO:0016265) and ”cell death” (GO:0008219). Apparently the student’s t model is too conservative in this case to provide interesting results.

To summarise the results gained by analysing three biological data sets, a student’s t model is always found to be preferred compared to a Gaussian model during the model selection process. Differences between the results of the t and normal distribution model are visible for the statistical inference results where the student’s t model tends to be more conservative w.r.t. general behaviour of the data set.

In order to check for differences between the biological implications of the 2 models, Gene Ontology analyses have been performed. The trend is the same as for the respective inference results. For high amounts of differentially expressed genes, the t model tends to find more GO terms than the normal model. Whereas for low amounts of such genes it provides a focus on certain terms which are also found to be significantly enriched by the Gaussian model.

8 Summary

The goal of this thesis was to test a possible approach of robustifying a hierarchical Bayesian ANOVA model designed for the analysis of microarray data. Additionally, it was meant to provide an understanding of the differences of results compared to the standard model and their differing biological implications. The specific aim was to gain robustness of the model likelihood function, i.e. the underlying error model. For this purpose a whole set of distributions was defined from which the likelihood function could be chosen.

Robust statistics are meant to work in a setting where not all assumptions of the standard - in our case Gaussian - model are fulfilled, but one is close enough to this setting except for some abnormal values. Thus certain properties are shared with the standard model, e.g. for an error model unimodal, symmetric distributions are reasonable. Additionally the distribution should be parametric and provide an up to a certain degree analytically tractable model. As the goal was specifically to deal with outlying values, the distributions were to have a high probability mass in their tails. Thus the usage of non-central student's t distributions was a reasonable choice.

For the actual calculations and statistical inference a MCMC algorithm has been designed and implemented. It made use of dimension changing moves which allow the algorithm to switch between different parameter spaces. These were not only required for modelling the biological setting (differential expression vs. no differential expression), but also to gain the possibility to include the standard (Gaussian) model into the model selection process. Aside from the implementation, some theoretical considerations of convergence behaviour were taken, as far as they could be deduced from the individual transition kernels. For practical handling of convergence assessment, auxiliary diagnostics had to be used, such as the ones provided by the R package coda.

Before applying the algorithm to microarray data, tests with artificial data sets were performed to assess its general behaviour. These test results met our expectations, as the algorithm was able to accurately and precisely estimate the underlying likelihood setting for all test data sets. Additionally it provided insights into required burn-in numbers and run lengths, as well as mixing problems for not well-chosen specifications of the underlying 'data set of possible likelihood functions'.

Applying the algorithm to three data sets of different biological settings has finally provided an answer to the question, if a student's t distribution was a reasonable model distribution for such data sets. For all three data sets student's t distributions with low degrees of freedom (4 or 10) were selected which significantly differ from a normal distribution. Differences in statistical inference results also led to different biological implications which showed the importance of handling the choice of a model likelihood with great care.

Table of mathematical expressions and functions

$id_R(.)$	identity of . in space R
$\mathcal{I}(\theta)$	Fisher's information (matrix)
$\mathbb{I}_A$	indicator function
E_n	n-dimensional unit matrix
$\mathcal{X}$	state space of observations
$\mathcal{B}(\mathcal{X})$	Borel sets on $\mathcal{X}$
Γ	class of distributions
$\delta_x(.)$	Dirac Delta function with mass in x
$supp_f$	support of f; i.e. the set of points on which f is nonzero
$\xi(.)$	target density
$\mathcal{K}(.,.)$	transition kernel

List of Figures

List of Tables

9 Bibliography

[1] Bioconductor, 2008.

[2] R project of statistical computing, 2008.

[3] Hein A., Richardson S., Causton H., Ambler G., and Green P. Bgx: a fully bayesian integrated approach to the analysis of affymetrix genechip data. *BMC Bioinformatics*, 2008.

[4] M. et al. Affara. Understanding endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal? *Philosophical Transactions of the Royal Society B*, 362:1469–1487, 2007.

[5] F. et al. Al-Shahrour. Fatigo: a web tool for finding significant association of gene ontology terms with groups of genes. *Bioinformatics*, 20, 2004.

[6] F. et al. Al-Shahrour. Babelomics: a system biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acid Research*, 2006.

[7] D. et al. Allison. *DNA microarrays and related genomics techniques*. Chapman & Hall/CRC, 2006.

[8] K. Bae and B. Mallick. Gene selection using a two-level hierarchical bayesian model. *Bioinformatics*, 2004.

[9] P. Baldi and A. Long. A bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 2001.

[10] James O. Berger. *Statistical decision theory*. Springer, 1980.

[11] James O. Berger. An overview of robust bayesian analysis. *Test*, 3:5–124, 1994.

[12] Jose Bernardo. *Bayesian theory*. Wiley, 1995.

[13] J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10:3–41, 1995.

[14] S. P. Brooks, P. Giudici, and G. O. Roberts. Efficient construction of reversible jump markov chain monte carlo proposal distributions. *Journal of the Royal Statistical Society. Series B*, 65:3–55, 2003.

[15] Stephen Brooks. Markov chain monte carlo method and its application. *The Statistician*, 47:69–100, 1998.

[16] Olivier Cappe, Christian P. Robert, and Tobias Ryden. Reversible jump, birth-and-death and more general continuous time markov chain monte carlo samplers. *Journal of the Royal Statistical Society. Series B (Methodological)*, 65:679–700, 2003.

[17] Siddharta Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49:327–335, 1995.

[18] Peter Congdon. *Bayesian Statistical Modelling.* Wiley, 2001.

[19] M. K. Cowles and B. P. Carlin. Markov chains monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91:883–904, 1996.

[20] D. Gamerman and H. Lopes. *Markov Cahin Monte Carlo, Stochastic Simulation for Bayesian Inference.* Chapman and Hall, 2006.

[21] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman and Hall, 1996a.

[22] D. et al. Glynn. David: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4, 2003.

[23] R. et al. Gottardo. Statistical analysis of microarray data: a bayesian approach. *Biostatistics*, 4:597–620, 2003.

[24] R. et al. Gottardo. Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics*, 62:10–18, 2006.

[25] Peter J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[26] J. et al. Ibrahim. Bayesian models for gene expression with dna microarray data. *J. Am. Stat. Assoc.*, 97:88–99, 2002.

[27] David Rios Insua and Fabrizio Ruggeri (Eds.). *Robust Bayesian Analysis.* Springer, 2000.

[28] H. Ishwaran and J. Rao. Detecting differentially expressed gene in microarrays using bayesian model selection. *J. Am. Stat. Assoc.*, 98:438–455, 2003.

[29] Friedrich Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physika Verlag, Heidelberg, Germany, 2002.

[30] A. Lewin, N. Bochkina, and S. Richardson. Fully bayesian mixture model for differential gene expression: Simulations and model checks. *Statistical Applications in Genetics and Molecular Biology*, 6, 2007.

[31] D. Lunn, N. Best, and J. Whittaker. Generic reversible jump mcmc using graphical models. *Technical Report,EPH-2005-01*, 2005.

[32] Mathworks. Matlab, 2004.

[33] R. et al. Nagoshi. The control of alternative splicing at genes regulating sexual differentiation in d. melanogaster. *Genome Biology*, 4, 2003.

[34] Jose C. Pinheiro, Chuanhai Liu, and Ying Nian Wu. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t-distribution. *Journal of Computational and Graphical Statistics*, 10:249–276, 2001.

[35] Sylvia Richardson and Peter J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59:731–792, 1997.

[36] Christian P. Robert. *The Bayesian choice.* Springer, 2001.

[37] Christian P. Robert. *Monte Carlo statistical methods.* Springer, 2004.

[38] G. O. Roberts and J. S. Rosenthal. Markov chain monte carlo: Some practical applications of theoretical results. *Can. J. Stat.*, 26:5–31, 1998.

[39] G. O. Roberts and J. S. Rosenthal. Two conbergence properties of hybrid samplers. *Ann. Appl. Prob.*, 8:397–407, 1998.

[40] G. O. Roberts and J. S. Rosenthal. Harris recurrence of metropolis-within-gibbs and trans-dimensional markov chains. *Ann. Appl. Prob.*, 16:2123–2139, 2006.

[41] J. et al. Shima. The murine testicular transcriptome: characterizing gene expression in the testis during the progression of spermatogenesis. *Biol. Reprod.*, 71:319–330, 2004.

[42] N. D. Shyamalkumar. Likelihood robustness. In David Rios Insua and Fabrizio Ruggeri, editors, *In Robust Bayesian Analysis.* Springer, 2000.

[43] Peter Sykacek, David P. Kreil, Lisa A. Meadows, Richard P. Auburn, Bettina Fischer, Steven Russell, and Gos Micklem. The impact of quantitative microarray optimization on gene expression analysis.

[44] Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Applied Statistics*, 22:1701–1762, 1994.

[45] Luke Tierney. A note on metropolis-hastings kernels for general state spaces. *The Annals of Applied Statistics*, 8:1–9, 1998.

[46] V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98:5116–5121, 2001.

[47] P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman and Hall, 1991.

[48] H. et al. Zhao. Multivariate hierarchical bayesian model for differential gene expression analysis in microarray experiments. *BMC Bioinformatics*, 2008.