Supplementary Material for "Biological Implication of Robust Noise Models"

Alexandra Posekany, Klaus Felsenstein and Peter Sykacek

Contents

1 Mathematical Structure of the model

1.1 Model structure

The Bayesian hierarchical model, which we use in our paper for investigating robustness, is based on a latent variable implementation of a biological indicator variable I_g. Furthermore it is linked with an ANOVA type linear model:

$$y_{n,g} = x_{n,g}^T \beta_g + \varepsilon_{n,g}, \qquad n = 1, \ldots, N, g = 1, \ldots, G \tag{1.1}$$

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; $x_{n,g} = [\mathbb{I}(S_{n,g} = 1), \ldots, \mathbb{I}(S_{n,g} = S)]^T \in \mathbb{R}^{S \times 1}$
$S_{n,g}$	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 $(x_{1,g}, \ldots, x_{N,g}) =: X_g = X \in \mathbb{R}^{S \times N}$ 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. Mathematically, this indicator differs between a univariate and a multivariate linear model by determining the dimension of β_g. Biologically it distinguishes between differential expression and non-differential expression of a gene. A one dimensional 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".

$$\begin{aligned} I_g = 0: \quad \beta_{g,0} | I_g = 0 \quad &\sim \quad N_1(\mu_{g,0}, (\tau_{g,0})^{-1}) \\ &\beta_g = [\beta_{g,0}, \ldots, \beta_{g,0}]^T \in \mathbb{R}^{S \times 1} \end{aligned} \tag{1.2}$$

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.

$$\begin{aligned} I_g = 1: \quad \beta_g | I_g = 1 \quad &\sim \quad N_S(\mu_g, T_g^{-1}) \\ &\mu_g = [\mu_{g,1}, \ldots, \mu_{g,S}]^T \in \mathbb{R}^{S\times 1} \\ &T_g = \begin{pmatrix} \tau_{g,1} & & 0 \\ & \ddots & \\ 0 & & \tau_{g,S} \end{pmatrix} \in \mathbb{R}^{S\times S} \end{aligned} \tag{1.3}$$

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.

$$I_g | p \quad \sim \quad Bin(1, p) \tag{1.4}$$

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

$$p \quad \sim \quad Be(a, b). \tag{1.5}$$

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.

1.2 Student's t Model

For the ansatz we take the general framework of the model described in section 1.1. Hereby the approach towards robustification is made using Student's t-distributions in the likelihood and prior setting. It is a well known (see [21]) and easily proved fact, 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:

$$X \sim t_\nu(\mu, \sigma^2) \Leftrightarrow \begin{array}{l} X|\varphi \sim N(\mu, \frac{1}{\varphi}\sigma^2) \\ \varphi \sim Ga(\frac{\nu}{2}, \frac{\nu}{2}) \end{array} \tag{1.6}$$

According to (1.6) the model is written as

$$\begin{aligned} y_{n,g}|\beta_g, \nu \quad &\sim \quad t_\nu(x_{n,g}^T \beta_g, \tau_\varepsilon)^{-1}) \quad \Leftrightarrow \\ y_{n,g}|\beta_g, \varphi \quad &\sim \quad N(x_{n,g}^T \beta_g, (\varphi_{n,g}\tau_\varepsilon)^{-1}) \\ \varphi_{n,g}|\nu \quad &\sim \quad Ga(\tfrac{\nu}{2}, \tfrac{\nu}{2}) \\ \tau_\varepsilon|g, h \quad &\sim \quad Ga(g, h) \end{aligned} \tag{1.7}$$

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 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.

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

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

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

□

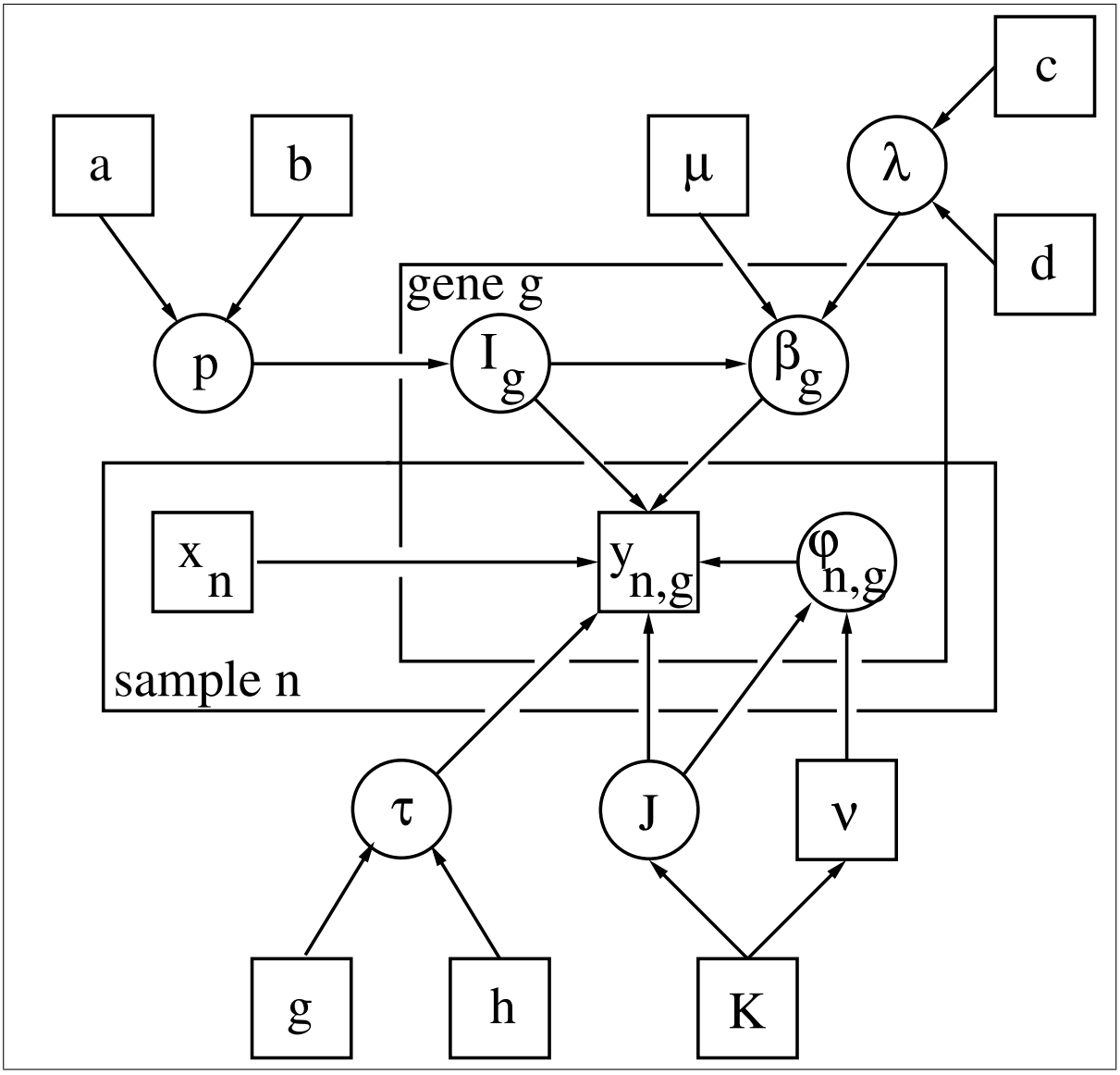


Figure 1.1: Directed Acyclic Graph representation of the model rectangular frames refer to variables which are fixed during the updates (data, fixed hyperparameters), variables in circles are updated as parts of the model

$y_{n,g}$	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
J	multinomial variable containing model probabilities

An essential component of the model in figure 1.1 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 hyper prior on the set $\mathfrak{N}$ over the parameter ν is specified:

$$\begin{aligned} \nu &\sim U_{\mathfrak{N}} && (1.8)\\ &\quad \mathfrak{N} := \{x \in \mathbb{R} | 2 \leq x := j \cdot c_{grid} \leq \nu_{max}, j \in \mathbb{N}\} && (1.9)\\ \Leftrightarrow &\quad \mathbb{P}[\nu = k|K] = 1/K, \;\; k \in \mathfrak{N}; \;\; K = |\mathfrak{N}| && (1.10) \end{aligned}$$

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 1.1.
However the definition of the set $\mathfrak{N}$ in equation (1.9) offers us great flexibility in the choice of the underlying parameter space and thus also for 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 multinomially distributed auxiliary variable J represents these model probabilities in a vector. Its update is truly equivalent to updating the model probabilities of all considered models together. The probability of choosing each model i m_i times among M draws is

$$\mathbb{P}[J = (m_1, \ldots, m_K)] = \frac{M}{m_1! \ldots m_k!} p_1^{m_1} \ldots p_K^{m_K}. \quad (1.11)$$

According to the above assumptions (Equation 1.8), the prior values of hyperparameters p_i equal $1/K$ for all i. Their updated posterior values are equal to the posterior probability of each model. The biological indicator for differential expression follows a Bernoulli distribution

$$\pi(I_g|p) \;\; = \;\; p^{I_g}(1-p)^{1-I_g} \quad (1.12)$$

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

$$\pi(p) \quad = \quad \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}. \tag{1.13}$$

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 1.1.
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 is described by one common parameter for all prior precision parameters and follows a Gamma distribution, i.e.

$$\tau_{g,s} \quad := \quad \lambda \quad \forall g, s \tag{1.14}$$

$$\lambda \quad \sim \quad Ga(c,d) \tag{1.15}$$

This reduces the model parts (1.2), (1.3) to:

$$\begin{array}{lll} I_g = 0 \quad \beta_{g,0}|I_g & \sim & N_1(\mu, (\lambda)^{-1}) \\ & & \beta_g = [\beta_{g,0}, \ldots, \beta_{g,0}]^T \ \in \mathbb{R}^{S\times 1} \\ I_g = 1 \quad \beta_g|I_g & \sim & N_S(\mu, (\lambda)^{-1} E_S) \end{array} \tag{1.16}$$

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

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

Table 1.1: Overview over Student's t model

1.3 Robustness

1.3.1 The Concept of 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 non informative priors can be seen as an example for cases, when a single type of distribution is selected with a specific goal in mind. Interpretation and computational 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 [39]:

> *The problem is not that Bayesians have yet to discover the truly non informative 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 two main problems that occur, when working with exponential family distributions (see [3]):

- 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.

1.3.2 Global Robustness

Because the concept of global robustness of priors will play a certain role in the focus of our work, the basic ideas of this theory will be presented here. The principal idea is to define a class of prior distributions Γ 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 [3]) The concept is rather vaguely defined and leaves it to the statistician to decide on the thresholds for "too large" as well as the quantity of interest.

$$\begin{aligned} r(\Gamma) &= \|\overline{\psi} - \underline{\psi}\|, \\ & \overline{\psi} = \sup_{\pi\in\Gamma} \psi(\pi, f), \quad \underline{\psi} = \inf_{\pi\in\Gamma} \psi(\pi, f), \end{aligned} \tag{1.17}$$

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

$$\Gamma' \subseteq \Gamma \Rightarrow (\overline{\psi}' - \underline{\psi}') \leq (\overline{\psi} - \underline{\psi}) \tag{1.18}$$

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.

1.3.3 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 they were in the focus of most criticism. Yet logically the likelihood function has considerable influence, as it determines the way how the data will influence the results. However, there is no easy way of quantifying the actual influence. This 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 ([32]) 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 (1.17)).

$$\overline{\psi} = \sup_{f \in \Gamma_f} \psi(\pi, f), \underline{\psi} = \inf_{f \in \Gamma_f} \psi(\pi, f), \tag{1.19}$$

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 adaptive nature 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 paper 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 [19]). For example Baldi et al. [2] 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. [16], Zhao et al. [42] and Gottardo et al. [12]. 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 [37]). 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 ([23]), rarely on t distributions ([13]). In some ways our modelling attempt is similar to Gottardo et al.'s ([13]), yet in other aspects 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.

1.3.4 Robustness as we see it

Different components of a probabilistic model can be aims for robustness considerations. Our main focus 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 [16], hierarchical model [42]; Bayesian ANOVA for microarrays [19]). 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 ([3]).

In the context of microarrays it has been used by [13]. 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 outlying observations. 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 approach focuses 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:

$$\Gamma = \{\{t_\nu(\mu, \tau^{-1}), \nu \in \mathfrak{N} \setminus \{\nu_{max}\}\}, N(\mu, \tau^{-1})\} \tag{1.20}$$

The definition of the set $\mathfrak{N}$ in (1.9) 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 [13]. However our goals differ, as we aim to compare 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.

2 MCMC schemes

2.1 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.

2.1.1 Monte Carlo integration

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

$$\mathbb{E}_f[h(X)] \quad = \quad \int_{\mathcal{X}} h(x)f(x)dx \tag{2.1}$$

Given the observations $(X_1, \ldots, X_n)$ which have been generated from the density $f(.)$ the expression (2.1) can be approximated by the *empirical average*

$$\overline{h_n} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} h(x_i) \tag{2.2}$$

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

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

Thus the term

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

2.1.2 Overview over used MCMC methods

This methodology is based on 2 principles. Firstly, it approximates expectations by averages of random draws from a given distribution (Monte Carlo). Secondly, it uses Markov chains which behave under certain conditions like draws from a single stationary distribution (see [31, 11] and [4]). With the MCMC algorithm we can estimate the means of our model parameters w. r. t. the posterior distribution.
As different variable settings require certain ways of sampling, the resulting hybrid sampler consists of Metropolis Hastings (MH), Gibbs and Reversible Jump (RJ) steps. Gibbs steps are the most easy to determine and implement. The principal idea of the Gibbs update is to draw each variable from the full conditional distribution given all other variables.For all variables with conjugate priors a direct update is possible, because we can explicitly determine the full conditional distribution of this model parameter given all others. In scenarios, where the Gibbs update is feasible, it is very effective, as every draw is automatically accepted.
The Gibbs scheme however cannot be applied to variables that do not have conjugate priors. For those parameters a more general updating method is used: the MH update. During a Metropolis Hastings step a new value for the parameter ϑ_{new} is proposed, i.e. it is drawn from a proposal distribution given the old parameter value $q(.|\vartheta_{old})$. For this proposed value an acceptance probability $\alpha(\vartheta_{old}, \vartheta_{new})$ is calculated, which compares the new to the old value in the following way:

$$\alpha(\vartheta_{old}, \vartheta_{new}) = \min\left\{\frac{\xi(\vartheta_{new}|\ldots)}{\xi(\vartheta_{old}|\ldots)}\frac{q(\vartheta_{old}|\vartheta_{new},\ldots)}{q(\vartheta_{new}|\vartheta_{old},\ldots)}, 1\right\}. \tag{2.3}$$

With this probability $\alpha(\vartheta_{old}, \vartheta_{new})$ the new value is accepted. The dependence on all other variables which stay unchanged during this step, is expressed by "...". As we use the Metropolis-Hastings update just for a part of the parameters (summarised by ϑ) at once, the target distribution $\xi(\theta|\ldots)$ is the marginal distribution of the possibly multivariate parameter θ given all other parameters $(\ldots)$. According to [4] this is a proper way of expressing a MH update in an hybrid sampler, because the other variables do not contribute to the dimensionality of the problem, as they remain fixed during the step. [4] have discussed the proper way of expressing MH steps for hybrid samplers.
We use MH steps for changing between student's t noise models with different degrees of freedom ν. A generalisation of the MH step for models of different dimension is the RJ step which has been introduced by [14] in order to move between spaces of different dimensions. With this special update the algorithm firstly can switch between the univariate space of β_g and the multivariate one, if the gene expression indicator changes between 2 steps. Our ansatz of the ANOVA model uses the RJMCMC methodology, whereas [13] have used an approach which is equivalent to the RJ move but can use a Gibbs step. Secondly, we can change between the highest degrees of freedom student's t distribution and the Gaussian distribution model of smaller dimension with a RJ update. In our cases there is only one move possible at each time. Therefore, the acceptance probability for such a move looks like $\alpha(\vartheta_{old}, \vartheta_{new})$ in equation (2.3) with an additional factor. This factor results from the Jacobian of a bijection between equally dimensional parameter spaces which the originals ones are embedded in.

2.2 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 he special case of ν_{max} results even in a reversible jump step (see below 2.2.2). 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 conditional distributions the common distribution of all modelled stochastic variables has to be derived:

$$p((I_g)_{g=1,\dots,G}, p, (\beta_g)_{g=1,\dots,G}, \nu, (\varphi_{n,g})_{n_1,\dots,N;g=1,\dots,G}, \tau_\varepsilon) \propto \tag{2.4}$$

$$\begin{aligned} \propto \quad & p(p|a,b)p(\tau_\varepsilon|g,h)p(J|K)p(\lambda|c,d) \\ & \textstyle\prod_g p(I_g|p)p(\beta_g|I_g,\mu_g,\tau_g) \\ & \textstyle\prod_n p(\varphi_{n,g}|\nu)p(y_{n,g} - x_{n,g}^T\beta_g|\varphi_{n,g}, I_g, \tau_\varepsilon) \end{aligned} \tag{2.5}$$

Changing dimensions of parameters and non-conjugate priors make the model complex and require a combination of Metropolis Hastings (MH), Gibbs (G) and Reversible Jump (RJ) steps. Algorithm 1 presents the structure of these steps using pseudo-code.

Algorithm 1 Hybrid MCMC Sampler

```
random initialisation of parameters
for n = 1 to burnin do
  c_grid = 1
  update parameters={
    update ν, J and φ_{n,g} jointly
    update p
    update λ
    update β_g and I_g jointly
    update τ }
end for
for n = 1 to burnin + simulationlength do
  c_grid = 0.05
  update parameters (see first burn-in)
end for
```

2.2.1 Initialisation & Choice of parameters

All parameters updated by the algorithm are initialised by random draw from their according prior distributions given the choice of parameters for the hyperpriors. The hyperparameters are chosen in line by keeping the underlying interpretation in mind (e.g. as prior counts etc.) and such as to avoid biasing inference.

- *Beta distribution*
  The hyperparameters of the conjugate Beta prior have a straight forward interpretation as number of genes, which have been observed to be (not) differentially expressed in a past experiment. A non informative choice would be $a = b = 1$ or equally the choice of parameters of the Jeffreys prior in such a case $a = b = \frac{1}{2}$, which also puts equal weight on both outcomes and has been chosen for the algorithm.
- *Gamma distribution*
  Again an uninformative prior for the scale parameter λ has to be found. For a precision parameter the Jeffreys prior and maximum entropy prior lead to equal results:

  $$\pi(\lambda) \propto \frac{1}{\lambda}$$

  This type of prior in a normal - conjugate gamma setting can be seen as a special type of gamma distribution, Ga(0,0). As is the case for many non informative priors it is improper. This does not have to be a serious obstacle for using it, as long as the resulting posterior is not improper and we do not run into difficulties with model selection (see the respective comments in the paper). For the parameters $\lambda, \tau_\varepsilon$, which have a gamma prior, it will be shown that almost surely the posterior will be proper. For the details of the structure of posterior parameters see below.
  Lemma 1 (Proper posterior).

  Given the non informative prior for the parameters $\theta = \lambda, \tau_\varepsilon$

  $$\theta \sim Ga(0,0) \quad \Leftrightarrow \quad \pi(\theta) \propto \frac{1}{\theta}$$

  the posterior will almost surely be proper in the given setting.

  Proof. – Case c=d=0: $c^* > 0 \Leftrightarrow G > 0 \vee i_1 > 0$
  As for any test the number of genes has to be positive, this condition is automatically fulfilled.

  $$d^* > 0 \Leftrightarrow \begin{array}{l} \varphi > 0 \\ \exists g : \beta_g - \mu \neq 0 \end{array}$$

  As φ is drawn from a gamma distribution, it will be positive. The set $\{\beta_g = \mu \quad \forall g\}$ is a null set, i.e. $\mathbb{P}[\beta_g - \mu = 0 \quad \forall g] = 0$, hence the condition is fulfilled almost surely.
  – Case g=h=0: $g^* > 0 \Leftrightarrow N > 0 \wedge G > 0$
  As for any test the number of total experiments and the genes should be

positive in order to make any sense, this condition is automatically fulfilled.

$$h^* > 0 \Leftrightarrow \begin{array}{l} \varphi > 0 \\ \exists (n,g) : y_{n,g} - x_{n,g}^T \beta_g \neq 0 \end{array}$$

In analogy to above the second condition is fulfilled almost surely, as for any experiment with expression values $y_{n,g}$ the set $\{y_{n,g} = x_{n,g}^T \beta_g \quad \forall n, g\}$ is a null set, i.e. $\mathbb{P}[y_{n,g} - x_{n,g}^T \beta_g = 0 \quad \forall n, g] = 0$.

□

- *Uniform distribution*
  The parameter ν is defined on a finite discrete parameter space, the set $\mathfrak{N}$. Choosing a uniform prior distribution is not only a legitimate choice of a proper non-informative prior, it also has the advantage of simplify the expression of the acceptance probability of the respective Metropolis-Hastings step.
  The choice of set $\mathfrak{N}$ and its maximum value ν_{max} is influenced not only by mathematical reasoning but also by computational practicability. As the acceptance probability of the Metropolis-Hastings update for ν would reach unreasonably high or low values for the huge, 'real' data sets, refining the grid used was a practical option for making jumps to nearby values more likely and thus help with the mixing of the chain. The choice of the value of ν_{max} was somewhat tricky, as test-runs with varying this value between values of $\{30, 45, 55, 65, 75, 100, 150\}$ have shown that on the one hand the 'rule of thumb' value for approximation by normal distribution, 30, has proven to be too small to allow for differing between t-model data of higher degrees of freedom (10-15) and normal data, whereas on the other hand high values like 75, 100, 150 let the parameter find a local mode between 45 and 65 instead of clearly pointing towards ν_{max}, even for normally distributed data. Thus a moderate value of 65 for higher values of c_{grid} (e.g. 1) and 45 for finer grids was chosen as ν_{max}, which has proven to be adequate with the test data and also performed well with the 'real' data sets.

2.2.2 Update ν

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.

As an additional feature the commonly used Gaussian model was taken into account as well. 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 auxiliary variables all equal one.

Acceptance probability

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

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 2.2.2)

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

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:

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

2.2.3 Update τ_ε

The error model is updated in the following way:

$$\tau_\varepsilon|\ldots \sim Ga(g + \frac{NG}{2}, h + \frac{1}{2}\sum_{n,g} \varphi_{n,g}(y_{n,g} - x^T_{n,g}\beta_g)^2)$$

2.2.4 Update β_g and I_g

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.

The hyperparameter λ will be updated in the following way:

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

(WD) update β_g conditional on all other variables

case $I_g = 0$

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

case $I_g = 1$

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

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

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

The auxiliary variable

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

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

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

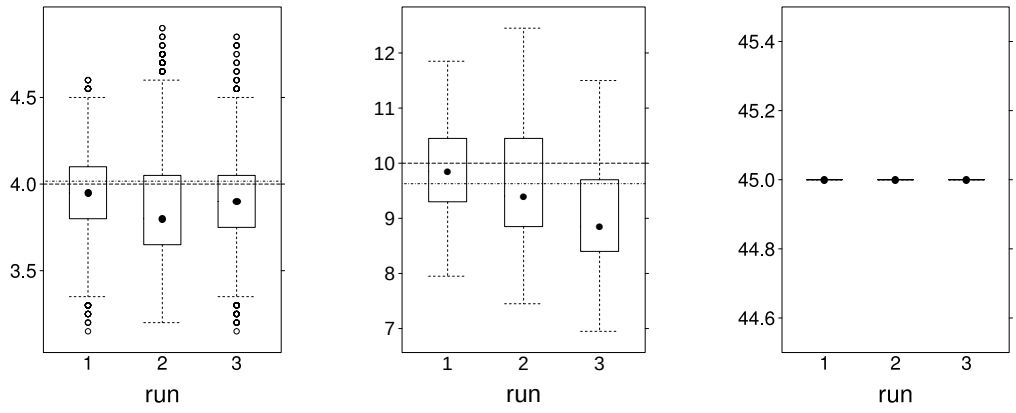
$$\begin{aligned}\beta_{g,0}|\varphi,\ldots &\sim N_1(\mu^*, (\varphi\lambda^*)^{-1}) \\ &\mu^* = \frac{\tau_\varepsilon \sum_{n=1}^{N} \varphi_{n,g} y_{n,g} + \lambda\mu}{\lambda^*} \\ &\lambda^* = (\tau_\varepsilon \sum_{n=1}^{N} \varphi_{n,g} + \lambda)\end{aligned}$$

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

3 Simulations

3.1 Inferring the noise on Artificial Data Sets

An important feature of the proposed algorithm which we tested for all synthetic data sets is to determine the underlying error distribution correctly, independently of the data's variance. Figure 3.1 summarises the distributions of the samples of ν as box plots. The sampler draws around the true value, while the variation, estimated by the interquartile range, is less than 2 degrees of freedom. In case of true Gaussian data the MCMC sampler stays with the Gaussian model without ever leaving it again. As the Gaussian fits the data very well, a move to the more complex 44 degrees of freedom t-distribution model is very unlikely. As the model identifies all error distributions correctly we may conclude that the proposed algorithm is well suited for identifying the required robustness level in real microarray data.



(i) t_4 data (ii) t_{10} data (iii) Gaussian data

Figure 3.1: The box plots represent the posterior distribution of the estimated degrees of freedom parameter for a t_4, a t_{10} and a Gaussian data set. The strong dashed line marks the true degrees of freedom value and the dash-dotted line marks the posterior mean of all three data sets per setting, i. e. 4.21, 10.66, $45 \sim \infty$

The simulations on artificial data also revealed that adjusting the grid size c_{grid} during runtime improves mixing and thus the convergence properties of the Markov chain. During the burn-in phase the grid size is refined from an initial value in the range of 1 to 5, as proposed in [13], to a smaller value of about 0.05 which stays fixed during

the following sampling process, see Algorithm 1. A relatively large grid size of about 1 allows the algorithm to quickly determine the approximately correct error model. Reducing the grid size after the first half of burn-in to $c_{grid} \approx 0.05$ improves mixing of the Markov chain without limiting the possibility of the algorithm to reach distant degrees of freedom. Defining the set ν flexibly allows us to infer the degrees of freedom ν via a discrete random variable J, as well as to approximate the continuous *true* degrees of freedom with high accuracy. Additionally, the MCMC sampler with the reduced grid size requires less updates to reliably infer the degrees of freedom.

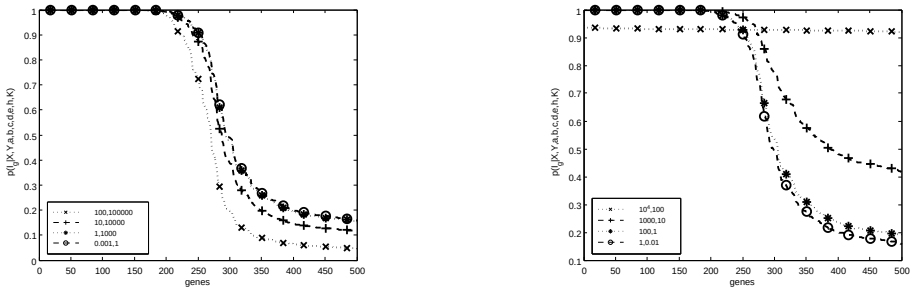
3.1.1 Sensitivity Analysis

This section provides further results regarding the models sensitivity to variations of hyper parameters. Following the arguments in the paper, we focus on an investigation of the hyper-parameters c and d which specify the hierarchical prior over ANOVA parameters β_g (*cf.* Figure 1.1). Sensitivity to variations of c and d is induced, if their effect in the marginal posterior $p(\lambda|c, d, X, Y)$ can not be neglected. This is the case if c and d are large in relation to the evidence contributed from the observations $X = \{x_n|\forall n\}$ and $Y = \{y_n|\forall n\}$. Figures 3.2 and 3.3 illustrate the results of these sensitivity analyses for Gaussian and Student-t distributed noise models (the degrees of freedom of the Student-t distribution were fixed to 4). We include the Gaussian case to illustrate that a potential sensitivity of the model to hyper parameter settings is not caused by our choice of a more involved noise model. The two subplots (i) shown in both figures are simulations with a prior expectation of λ, $E[\lambda]_{p(\lambda|c,d)} = \frac{c}{d}$ being 0.001. Large (c, d) values lead in this case to using a zero mean Gaussian prior over ANOVA parameters β_g with a variance of 1000. This has the consequence that differentially expressed genes are hard to detect. The two subplots (ii) shown in both figures are simulations with a prior expectation of λ, $E[\lambda]_{p(\lambda|c,d)} = \frac{c}{d}$ being 100. Large (c, d) values lead in this case to using a zero mean Gaussian prior over ANOVA parameters β_g with a variance of 0.01. This has ultimately the (meaningless) consequence that all genes are assessed as differentially expressed.

The additional sensitivity analyses provided here corroborate thus the results in the main paper that a hierarchical specification of the prior over ANOVA coefficients, β_g, with a Jeffreys prior $p(\lambda|c, d)$ is essential for warranting that the posterior probabilities $P(I_g|X, Y, a, b, c, d, e, h, K)$ are data driven and insensitive to local perturbations of the hyper-parameters c and d.

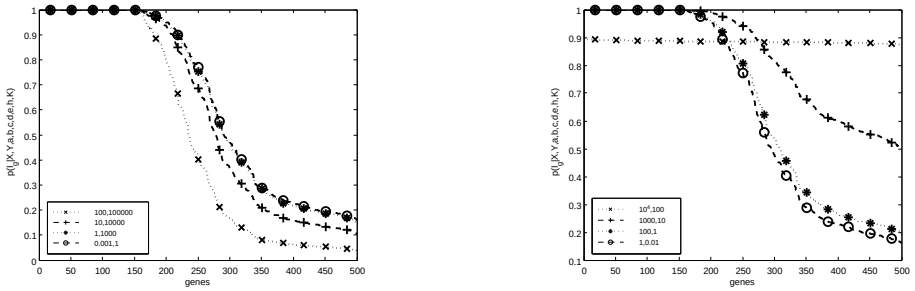
3.2 Biological Data Sets

To highlight the importance of choosing valid noise models for microarray analysis we applied the proposed inference scheme to fourteen microarray data sets, summarised in Table 3.1. We inferred differentially expressed genes for every data set from the Gaussian and the estimated optimal Student-t model to obtain a quantitative statement. This approach resulted in two lists of differentially expressed genes with the intersect



(i) (ii)

Figure 3.2: Gaussian noise model based sensitivity analysis with different (c, d) parameter settings. Both subplots illustrate the dependency of the ordered posterior probabilities for differential expression $P(I_g|X, Y, a, b, c, d, e, h, K)$ on hyper parameter settings. Inference was based on the synthetic experiment which is described in the paper. Subplot (i) uses a parametrisation of (c, d) such that the prior expectation of λ, $E[\lambda]_{p(\lambda|c,d)} = \frac{c}{d}$ equals 0.001. Subplot (ii) sets (c, d) such that the prior expectation of λ, $E[\lambda]_{p(\lambda|c,d)} = \frac{c}{d}$ equals 100. By reducing the prior variance $V[\lambda]_{p(\lambda|c,d)} = \frac{c}{d^2}$, the hyper parameters get in both situations increasingly informative and start affecting the $P(I_g|X, Y, a, b, c, d, e, h, K)$ values. Informative (c, d) combinations lead in subplot (i) to smaller posterior probabilities for differential expression, whereas subplot (ii) shows for informative (c, d) combinations larger posterior probabilities of differential expression. This behaviour is in line with theoretical expectations and demonstrates that the hierarchical prior together with uninformative choices for (c, d) is essential for getting data driven results.



(i) (ii)

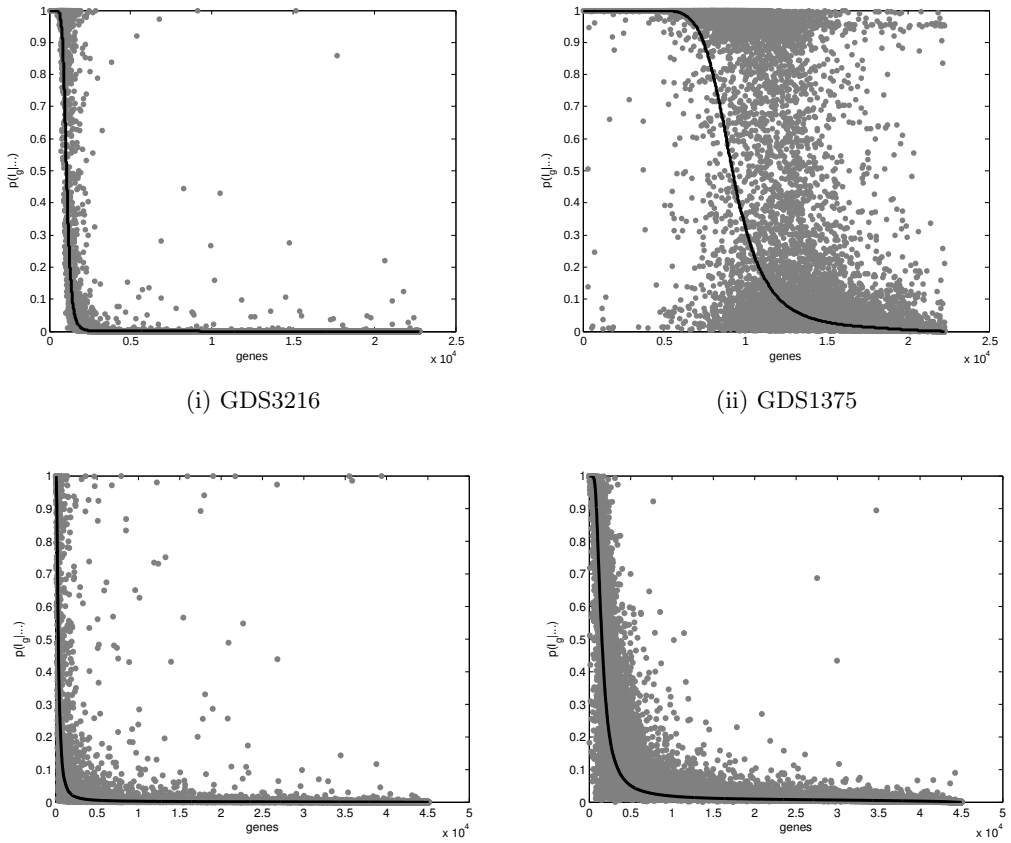
Figure 3.3: A sensitivity analysis with different (c, d) parameter settings lead for the Student-t noise model to the same effect that was observed in Figure 3.2 when using Gaussian noise. Both subplots illustrate the dependency of the ordered posterior probabilities for differential expression $P(I_g|X, Y, a, b, c, d, e, h, K)$ on hyper parameter settings. Inference was based on the synthetic experiment which is described in the paper. Subplot (i) uses a parametrisation of (c, d) such that the prior expectation of λ, $E[\lambda]_{p(\lambda|c,d)} = \frac{c}{d}$ equals 0.001. Subplot (ii) sets (c, d) such that the prior expectation of λ, $E[\lambda]_{p(\lambda|c,d)} = \frac{c}{d}$ equals 100. By reducing the prior variance $V[\lambda]_{p(\lambda|c,d)} = \frac{c}{d^2}$, the hyper parameters get in both situations increasingly informative and start affecting the $P(I_g|X, Y, a, b, c, d, e, h, K)$ values. Informative (c, d) combinations lead in subplot (i) to smaller posterior probabilities for differential expression, whereas subplot (ii) shows for informative (c, d) combinations larger posterior probabilities of differential expression. This behaviour is in line with theoretical expectations and demonstrates that the hierarchical prior together with uninformative choices for (c, d) is essential for getting data driven results.

representing agreement and the symmetric difference representing different biological interpretations induced by an inappropriate noise model.

Table 3.2 contains the findings for the alternative normalisations and non-parametric methods. The arresting result of our evaluation is that a heavy-tailed Student-t noise model is a better fit than a Gaussian noise model for every considered data set independently of the normalisation. For most data sets a t-distribution with degrees of freedom between 1 and 5 got the highest posterior probability. This indicates the need for robust noise models which can handle outlying data points well and allows us to conclude that Gaussian noise models are unsuitable for microarray analysis, even if according to [29] only about 5 to 15 percent of samples are non-normally distributed.

The robust model is generally less sensitive to outlying values, as they appear to be closer to the bulk of the data. Models with t-distributed noise will therefore assign lower posterior probabilities of differential expression, if the classification is drawn by one or a few outlying values. In general outlying observations increase variance. Where outliers additionally lead to a decreased difference between average expression values, the Gaussian noise model will overlook differentially expressed genes which would be captured by heavier-tailed noise model. We therefore expect that a wrongly chosen noise model will lead to false positives and false negatives. This expectation is confirmed by the graphs in Figure 3.4 which illustrate such noise model dependencies of the posterior probabilities of differential expression for two of the datasets. This statement holds true independently of the applied normalisation method, as we can see from Figures 3.4, 3.5 and 3.6. Figure 3.4 shows the graphs for vsn normalised data, Figure 3.5 for the loess normalised data and 3.6 the quantile normalised data, respectively.

Each graph in Figure 3.4 is ranked w.r.t. the posterior probabilities obtained by setting the noise distribution either to a Gaussian density or the most probable t-distribution. The corresponding probabilities are shown as a decreasing curve. The probabilities which result from the other noise model are shown as grey dots. In these data sets we find both false positives and false negatives. On one hand several of the genes that have been considered highly differentially expressed by the Gaussian model have a much lower posterior probability in the robust model. On the other hand single genes or whole 'clusters' of genes which have low posterior probability for the Gaussian model are actually highly differentially expressed in the Student-t model. The human melanoma (GDS1375) data set, as seen in Figure 3.4 (b), is a good example for the appearance of a large cluster of such genes at the top right. Since the model inference over degrees of freedom ν clearly favours the robust Student-t model we can consider these genes as those which would have been overlooked in a normal distribution based model. Table 3.1 shows that the number of genes which show a noise model dependency in the assessment about differential expression range from 119 to 3561. This is about one tenth to two times the number of genes which are independently of the noise model assessed as differentially expressed. We can thus conclude that the choice of noise model can be very influential on inferred gene lists.



(i) GDS3216

(ii) GDS1375

(iii) GDS1555

(iv) GDS3162

Figure 3.4: Noise model dependent difference in posterior probability of differential expression. The graph in subplot (i) is ranked by the posterior probability of differential expression obtained with the most probable t-distributed noise model (probabilities shown as black line). The corresponding posterior probabilities obtained by using a Gaussian noise model are shown as grey dots. The graph in subplot (ii) is ranked by the posterior probability of differential expression obtained with a Gaussian noise model (probabilities shown as black line). The corresponding posterior probabilities obtained when using the most probable t-distributed noise model are shown as grey dots.

Org.	GEO ID	Reference	Prep.	N	$\overline{\nu}$	comm. genes	diff. genes	comm. GO terms	diff. GO terms
A. thal.	GDS3216	([9])	MAS5.0	12	4.71	1176	150	111	78
A. thal.	GDS3225	([38])	MAS5.0	4	5.50	832	290	161	21
D. rerio	GDS1404	([6])	PathStat	10	13.58	1776	136	11	14
D. mel.	GDS1686 (I)	([43])	RMA	9	3.62	136	174	11	96
H. sap.	CAMDA 08	([1])	CLSS4.1	24	4.04	400	304	26	67
H. sap.	GDS1375	([36])	MAS5.0	70	3.25	6861	3561	160	316
H. sap.	GDS810	([5])	MAS5.0	31	4.37	72	135	9	51
H. sap.	GDS2960	([41])	RGP3.0	101	4.33	318	166	51	2
M. musc.	GDS660	([33])	MAS5.0	22	10.48	584	126	20	26
M. musc.	GDS3221	([34])	RMA	24	4.21	180	119	108	52
M. musc.	GDS3162	([35])	MAS5.0	10	4.38	797	446	112	66
M. musc.	GDS1555	([27])	MAS5.0	8	3.90	131	183	24	110
R. nor.	GDS2946	([24])	MAS5.0	15	4.57	146	157	14	306
R. nor.	GDS972	([20])	MAS5.0	44	4.98	369	163	94	71
D. mel.	"Spike In"	([7])	MAS5.0	6	3.74	401	1748	-	-

Table 3.1: Overview of the biological data sets describing the organism (Org.), the GEO ID (CAMDA 08 refers to the Endothelial Apoptosis contest datasets of the meeting and "Spike In" to the "Golden Spike" experiment), the preprocessing method (Prep.), the overall number of arrays (N), the average degrees of freedom ($\overline{\nu}$), the number of common genes (comm.), the number of genes with noise model depending differential expression assessment (diff.), the number of common GO terms (comm.) and finally the number of noise model dependent GO terms (diff.). The GEO entry GDS1686 (I) refers to the behavioural subset of the data (only the sleep deprived flies). In column prep. we use MAS5.0 to refer to the Affymetrix MAS 5.0 quantisation method, RMA to refer to the "Robust Multi-array Average" method by [18] (both used for Affymetrix arrays), PathStat for referring to the package described in [28], CLSS4.1 to refer to the Codelink Software Suite 4.1 and RGP3.0 to refer to Research Genetics' Pathway software v. 3.0.

3.3 Alternative Normalisation

The results presented in the paper are based on data which has been normalised with vsn methodology (see [15]). To assure that the found effects are not due to this specific normalisation method, we have applied rma normalisation ([17]) to some of the data sets where CEL files were available. Furthermore, we selected a subset of data sets and applied loess and quantile normalisation as well as Liu's normalisation based on probe-level measurement error ([25],[26]). Then we compared the results to the ones of vsn normalised data.

We have listed the detailed results in Table 3.2. Again all data sets prefer a student's t model with low degrees of freedom. In case of loess and quantile normalised data, where variance stabilisation is apparently more important, the degrees of freedom estimate is even lower than for the vsn data. We compare the results for the example above, the human melanoma data (GDS1375). Here a student's t model with about 1.1 degrees of freedom has the highest posterior probability for both loess and quantile normalised data. Figure 3.4 shows the results for two of the data sets for which we have included graphs into the main paper. For the human melanoma data set the differences in the posterior probabilities are eye-catching, as a large percentage of genes is classified differently w. r. t. differential expression. The Arabidopsis data set represents the more typical case that differences are less obvious when looking at the gene lists. But as we showed in the main paper they gain weight when follow-up analyses, like in our case the Gene Ontology

analysis, are applied.

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

Table 3.2: Subset of data sets for testing alternative normalisations. We show the posterior mean degrees of freedom $\overline{\nu}$ and the numbers of common ('comm.') and different ('diff.') genes which the two methods classified as differentially expressed with a probability of more than 85%. In all cases t distributions with small degrees of freedom between 1 and 3 are preferred.

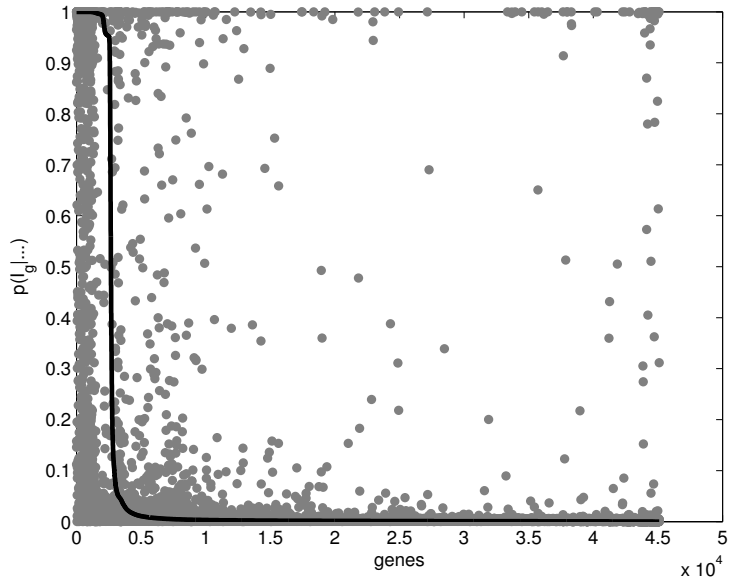
Interestingly we find that for loess or quantile normalised data the degrees of freedom of the selected optimal t model are in general much lower than for vsn normalised data. In almost all cases Cauchy-like t distributions were the preferred posterior model. This observation is consistent with the findings by [30]. A possible interpretation for this behaviour could be that these normalisation methods enforce outlying or non-Gaussian data points which are better explained by very heavy-tailed models.

3.3.1 Spike-in data

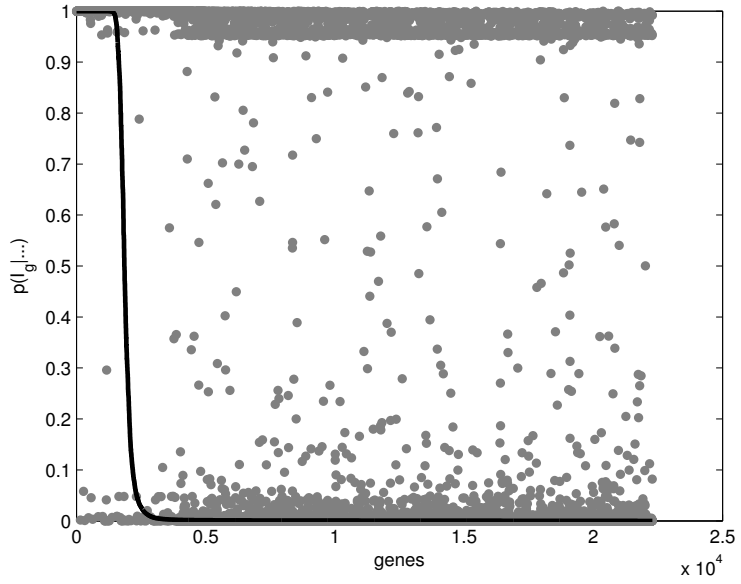
We wish to include variation into the model which may come from 2 different origins. Most importantly, variation of the biological components has to be included to reasonably perform inference on biological data. The laboratory work always adds noise to the measurements, to what amount however often remains unknown. We use the "Golden Spike" Experiment by [7] as a regularised data set where variation mainly stems from technical components. For this experiment sets of RNA were chosen to have a predefined fold change. The RNA was then hybridised to Affymetrix chips (for details see [7])

Firstly, we wanted to check the performance of our algorithm. Therefor we used the preprocessing applied by [7] When comparing it against the methods [7] have considered in their paper we found that our algorithm could compete with those methods w.r.t. sensitivity and false discovery rate at the top end.

Secondly, we applied our algorithm to identify the appropriate noise model. To make this data comparable to the majority of our data sets we used MAS5.0, as well as vsn for preprocessing. Our astounding results showed that, also for data where the main source of error is the laboratory work, a student's t model fits much better than a Gaussian one.

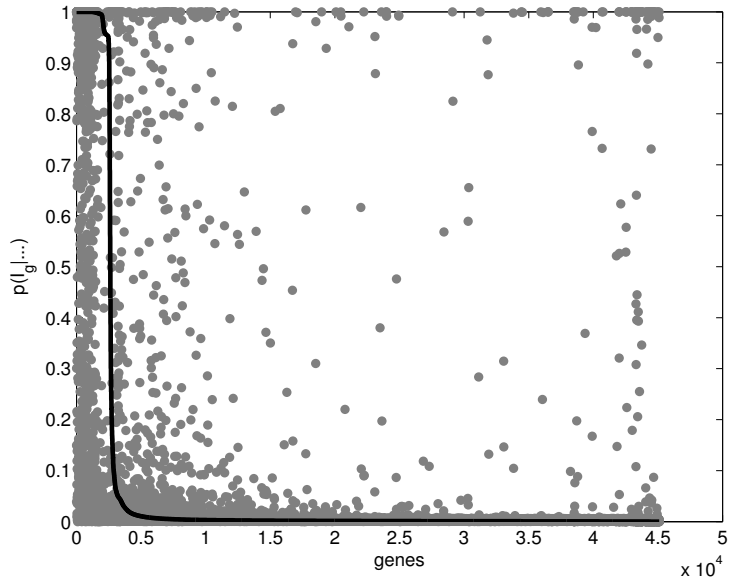


(i) GDS1555

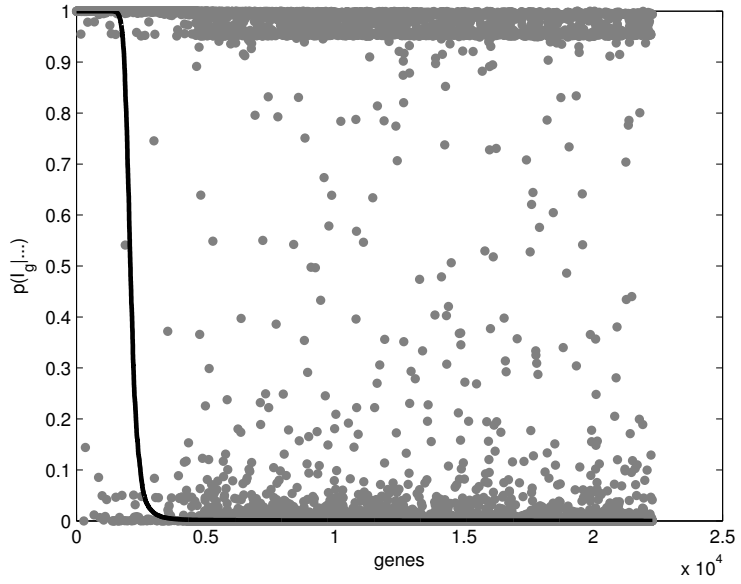


(ii) GDS1375

Figure 3.5: Difference in the ranked posterior probability of differential expression for loess normalised data; each graph is ranked separately the genes on the x axis are ordered w. r. t. decreasing posterior probability in the Gaussian model



(i) GDS1555



(ii) GDS1375

Figure 3.6: Difference in the ranked posterior probability of differential expression for quantile normalised data; each graph is ranked separately the genes on the x axis are ordered w. r. t. decreasing posterior probability in the Gaussian model

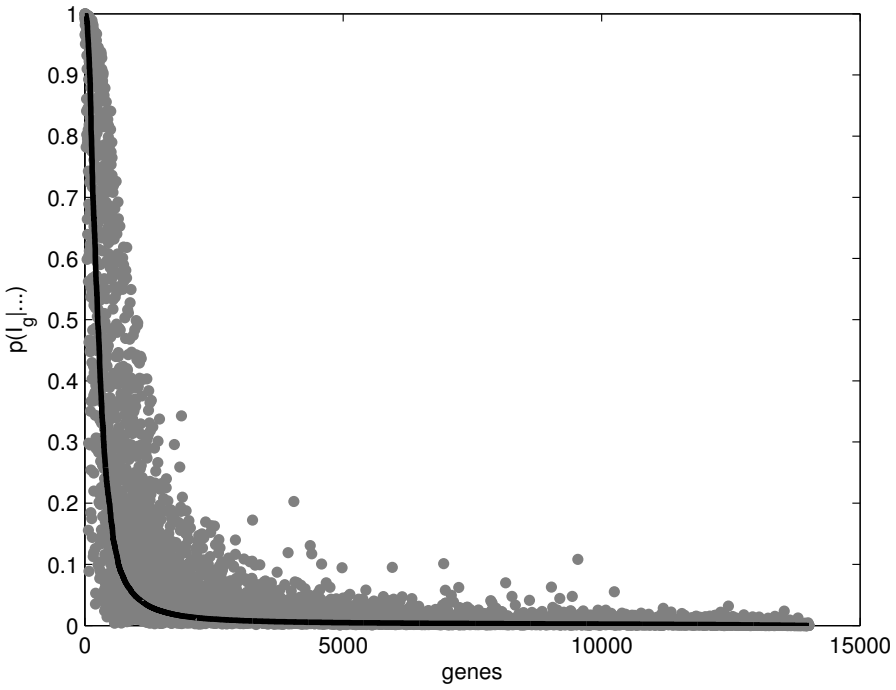


Figure 3.7: Difference in the ranked posterior probability of differential expression for spike-in data, normalised with vsn; each graph is ranked separately the genes on the x axis are ordered w. r. t. decreasing posterior probability in the Gaussian model

We compared the gene lists for genes with posterior probability of differential expression against the list of genes in [7] which have a fold-change rate greater than 2 according to the experimental settings. We could see see that the Student's t model showed a higher accuracy than the Gaussian model. At our chosen cut-off of 0.85 the Student's t model has an accuracy of 78% compared to 72% for the Gaussian model. Figure 3.8 plots the cut-off vs. the accuracy of the Gaussian model and the Student's t model on the spike-in data.

These results give us an indication that at least partially the non-normal behaviour of microarray data might be introduced by laboratory processes. However, there is no reason to believe that laboratory work is the only cause for heavy-tailed behaviour of the distribution of microarray data.

3.3.2 Non-parametric methods

Non-parametric methods are generally applied, when the validity of distribution assumptions is unknown and doubted. Such approaches are therefore commonly used for

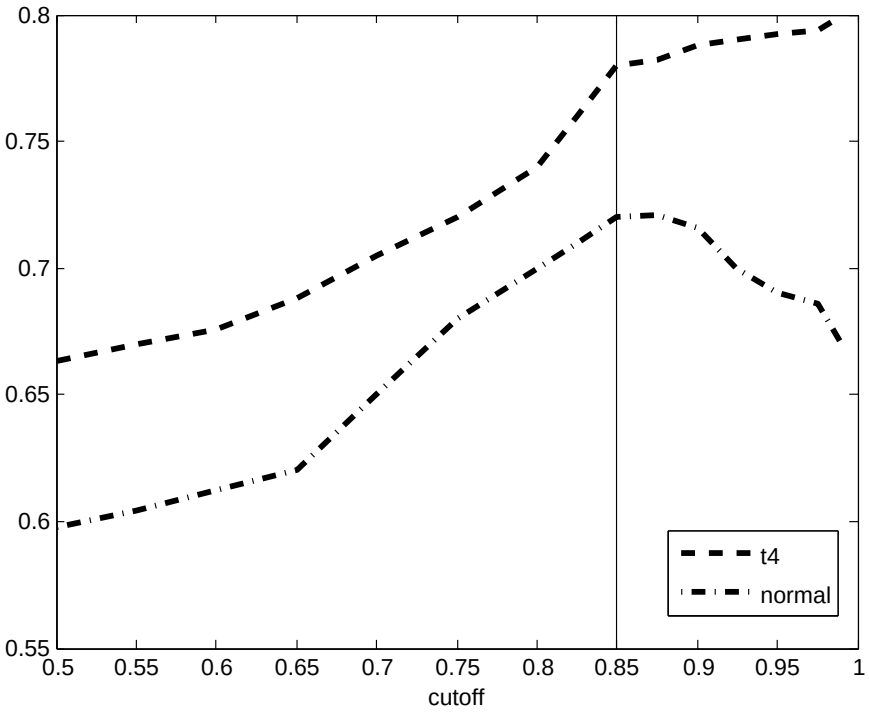


Figure 3.8: Plot of cut-off vs. accuracy. The Gaussian model has its optimal accuracy at at cut-off 0.85, which was for this reason empirically chosen as cut-off level for comparing the respective gene lists; then the accuracy is about 72%. Contrary to that, the Student's t model has its optimum of 80% at the highest cut-off of 0.99. At the level of 0.85 its accuracy is reasonably high at 78%.

robust assessment of microarray data, see [37] or [10]. We chose the following non-parametric methods for analysing microarray data: the Kruskal-Wallis test, the classical non-parametric version of one-way ANOVA, ANOVA on rank transformed and aligned rank transformed data, as described by [8], as well as permutation tests based on such (non-)parametric statistics, for example see [22].

Selecting these methods was motivated by good comparability, because they are non-parametric generalisations of a one-way ANOVA approach. The Kruskal-Wallis test is the non-parametric generalisation of the t test on ranked statistics. However, an approximately parametric distribution of its test statistic is assumed. To avoid this assumption a permutation test can instead be performed with the Kruskal-Wallis test statistic. We performed such a permutation test using 10000 permutations to estimate the distribution of the test statistic over the data set.

[8] evaluated several approaches for increasing the robustness of ANOVA models, including rank transforming the data as well as using robust mean estimates, for example like the truncated mean and the median. Since we wanted a non-parametric approach

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

Table 3.3: Subset of data sets for testing non-parametric methods. "KW perm" contains the fraction of shared results at a p-value cut-off of 1% for the Kruskal Wallis permutation test with the robust student's t and the Gaussian model, "RANOVA" the fraction of shared results for aligned rank transformed ANOVA with the robust student's t and the Gaussian model, respectively. A dash signifies that the non-parametric method could classify no gene as differentially expressed with a p-value smaller than 0.01 due to too small replicate or sample size.

towards ANOVA, we chose the ANOVA on (aligned) rank transformed data. In the one-way ANOVA setting, such as the one considered by us, there is no difference between the different approaches of rank transforming the data. Differences would only occur for interaction terms which assume that more than 1 factor is available and considered in the analysis. The results for both methods are listed in Table 3.3.

To compare the approaches we selected all genes with a p-value of differential expression above 1%. We then took the same number of top-ranked genes for the robust and the Gaussian model and calculated the relative amount of shared genes which were classified as differentially expressed. In cases with large enough sample and replicate size, the non-parametric methods generally share a slightly larger fraction of genes with the robust student's t model than the Gaussian one. The better agreement of robust methods shows that non-parametric approaches are in general to be preferred to parametric ones, if the given sample size allows to apply them. However, our analysis also revealed a major drawback of non-parametric approaches. Consistent with [40] we found that the non-parametric methods suffered from lack of power, when few samples or replicates were available. In cases, where only 2-3 replicates per group and 4-24 samples overall were provided, the non-parametric methods were unable to identify any significant genes (marked by the dashes in Table 3.3).

3.3.3 Probe-level measurement error

[25] have chosen a different approach towards robustifying their analysis by integrating effects on probe-level into their probabilistic model. They use this probabilistic normalisation approach to estimate the required variables for calculating the probe-level measurement error. In [26] they fit Gaussian kernels with variance components depending on the variation of probe-level measurements. To assess the validity of Gaussian

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

Table 3.4: The results for the mmgMOS and the PPLR method separately for the 3 data sets where we had CEL files available. 'comm.' and 'diff.' again signifies the number of common and different genes which are classified as differentially expressed by the robust and the normal model.

model assumptions for that kind of data we apply our algorithm to the posterior mean estimates of their model.

For testing whether such representations are an alternative to heavy tailed noise models, we applied our algorithm to multi-mgMOS normalised data. We also used it on the posterior expression estimates, obtained by the PPLR method, to test the model's Gaussian noise assumption. When applying the algorithm to the mmgMOS normalised data we found that the over all noise of the expressions followed a t distribution with degrees of freedom between 2 and 3, see Table 3.4 However, when analysing the PPLR model's expression estimates, they are heavier tailed than their mmgmos normalised input data, even though the model, inferring them, assumes Gaussian distributions, as shown in Table 3.4.

4 Bibliography

[1] M. Affara, Dunmore B., C. Savoie, S. Imoto, Y. Tamada, H. Araki, D. Charnock-Jones, S. Miyano, and C. Print. Understanding endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal? *Philosophical Transactions of the Royal Society B*, 362:1469–1487, 2007.

[2] 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*, 17:509–519, 2001.

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

[4] J. Besag et al. Bayesian computation and stochastic systems. *Statistical Science*, 10:3–41, 1995.

[5] E. Blalock, J. Geddes, K. Chen, N. Porter, W. Markesbery, and P. Landfield. Incipient alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences*, 101(7):2173–8, 2004.

[6] D. Cameron, K. Gentile, F. Middleton, and P. Yurco. Gene expression profiles of intact and regenerating zebrafish retina. *Molecular Vision*, 11:775–91, 2005.

[7] S. Choe, M. Boutros, A. Michelson, G. Church, and M. Halfon. Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biology*, 6:R16, 2005.

[8] J. De Haan, S. Bauerschmidt, R. van Schaik, E. Piek, L. Buydens, and R. Wehrens. Robust anova for microarray data. *Chemometrics and intelligent laboratory systems*, 98:38–44, 2009.

[9] J. Dinneny, T. Long, J. Wang, J. Jung, D. Mace, S. Pointer, C. Barron, S. Brady, J. Schiefelbein, and P. Benfey. Cell identity mediates the response of arabidopsis roots to abiotic stress. *Science*, 320(5878):942–5, 2008.

[10] X. Gao and P. Song. Nonparametric tests for differential gene expression and interaction effects in multi-factorial microarray experiments. *BMC Bioinformatics*, 6:186, 2005.

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

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

[13] R. Gottardo, A. Raftery, K. Yeung, and R. Bumgarner. Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics*, 62:10–18, 2006.

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

[15] W. Huber, A. Heydebreck, S. Sueltmann, A. Poustka, and M. Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2(1), 2003.

[16] J. Ibrahim, M-H. Chen, and R. Gray. Bayesian models for gene expression with dna microarray data. *J. Am. Stat. Assoc.*, 97:88–99, 2002.

[17] R. Irizarry et al. Summaries of affymetrix genechip probe level data. *Bioinformatics*, 31:e15, 2003.

[18] R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, and T. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 31:249–264, 2003.

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

[20] J. Jin, R. Almon, D. DuBois, and W. Jusko. Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays. *Journal of Pharmalcology and experimental therapeutics*, 307(1):93–109, 2003.

[21] N. Johnson and B. Welch. Applications of the noncentral t distribution. *Biometrika*, 31:362–389, 1940.

[22] M. Lee, G. Whitmore, H. Björkbacka, and M. Freeman. Nonparametric methods for microarray data based on exchangeability and borrowed power. *Journal of Biopharmaceutical Statistics*, 15:783–797, 2005.

[23] 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(1), 2007.

[24] S. Li, H. Zhang, C. Hu, F. Lawrence, K. Gallagher, A. Surapaneni, S. Estrem, J. Calley, G. Varga, E. Dow, and Chen Y. Assessment of diet-induced obese rats as an obesity model by comparative functional genomics. *Obesity (Silver Spring)*, 16(4):811–8, 2008.

[25] X. Liu, N. Milo, M.and Lawrence, and M. Rattray. A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644, 2005.

[26] X. Liu, N. Milo, M.and Lawrence, and M. Rattray. Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*, 22(17):2107–2113, 2006.

[27] N. MacLennan, L. Rahib, C. Shin, Z. Fang, S. Horvath, J. Dean, J. Liao, E. McCabe, and Dipple K. Targeted disruption of glycerol kinase gene in mice: expression analysis in liver shows alterations in network partners related to glycerol kinase activity. *Human Molecular Genetics*, 15(3):405–15, 2006.

[28] F. Middleton, E. Ramos, Y. Xu, H. Diab, X. Zhao, U. N. Dias, and M. Meguid. Application of genomic technologies: DNA microarrays and metabolic profiling of obesity in the hypothalamus and in subcutaneous fat. *Nutrition*, 20(1):14–25, 2004.

[29] J. Novak, S. Kim, J. Xu, O. Modlich, D. Volsky, D. Honys, J. Slonczewski, D. Bell, F. Blattner, E. Blumwald, M. Boerma, M. Cosio, Z. Gatalica, M. Hajduch, J. Hidalgo, R. McInnes, M. Miller, M. Penkowa, M. Rolph, J. Sottosanto, R. St-Arnaud, M. Szego, D. Twell, and C. Wang. Generalization of DNA microarray dispersion properties: microarray equivalent of t-distribution. *Biology Direct*, 1(1):27, 2006.

[30] E. Purdom and S. Holmes. Error distribution for gene expression data. *Statisitical Applications in Genetics and Molecular Biology*, 4(1):Article16, 2005.

[31] Christian P. Robert and Robert Casella. *Monte Carlo statistical methods*. Springer-Verlag, New York, 2004.

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

[33] C. Small, J. Shima, M. Uzumc, M. Skinner, and M. Griswold. Profiling gene expression during the differentiation and development of the murine embryonic gonad. *Biol. Reprod.*, 72(2):492–501, 2005.

[34] M. Somel, H. Creely, H. Franz, U. Mueller, M. Lachmann, P. Khaitovich, and S. Paabo. Human and chimpanzee gene expression differences replicated in mice fed different diets. *PLoS One*, 3(3):e1504, 2008.

[35] S. Someya, T. Yamasoba, G. Kujoth, T. Pugh, R. Weindruch, M. Tanokura, and T. Prolla. The role of mtdna mutations in the pathogenesis of age-related hearing loss in mice carrying a mutator dna polymerase gamma. *Neurobiological Aging*, 29(7):1080–92, 2008.

[36] D. Talantov, A. Mazumder, J. Yu, T. Briggs, Y. Jiang, J. Backus, D. Atkins, and Y. Wang. Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clin. Cancer Res.*, 11(20):7234–42, 2005.

[37] 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(9):5116–5121, 2001.

[38] D. Van Hoewyk, H. Takahashi, E. Inoue, A. Hess, M. Tamaoki, and E. Pilon-Smits. Transcriptome analyses give insights into selenium-stress responses and selenium tolerance mechanisms in arabidopsis. *Physiol. Plant.*, 132(2):236–53, 2008.

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

[40] E. Whitley and J. Ball. Statistics review 6: Nonparametric methods. *Critical Care*, 6(6):509–513, 2002.

[41] Z. Yao, J. Jaeger, W. Ruzzo, C. Morale, M. Emond, U. Francke, D. Milewicz, S. Schwartz, and E. Mulvihill. A Marfan syndrome gene expression phenotype in cultured skin fibroblasts. *BMC Genomics*, 8(1):319, 2007.

[42] H. Zhao, K. Chan, L. Cheng, and H. Yan. Multivariate hierarchical Bayesian model for differential gene expression analysis in microarray experiments. *BMC Bioinformatics*, 9(Suppl 1):S9, 2008.

[43] J. Zimmerman, W. Rizzo, K Shockley, D. Raizen, N. Naidoo, M. Mackiewicz, G. Churchill, and A. Pack. Multiple mechanisms limit the duration of wakefulness in Drosophila brain. *Physiol. Genomics*, 27(3):337–50, 2006.