Fully Bayesian mixture model for differential gene expression: simulations and model checks

Alex Lewin, Natalia Bochkina and Sylvia Richardson

Centre for Biostatistics, Department of Epidemiology and Public Health,
Imperial College, Norfolk Place, London W2 1PG, UK

January 12, 2007

Summary

We present a Bayesian hierarchical model for detecting differentially expressed genes using a mixture prior on the parameters representing differential effects. We formulate an easily interpretable 3-component mixture to classify genes as over-expressed, under-expressed and non-differentially expressed, and model gene variances exchangeably to allow for variability between genes. We show how the proportion of differentially expressed genes, and the mixture parameters, can be estimated in a fully Bayesian way, extending previous approaches where this proportion was fixed and empirically estimated. Good estimates of the false discovery rates are also obtained.

Different parametric families for the mixture components can lead to quite different classifications of genes for a given data set. Using Affymetrix data from a knock out and wildtype mice experiment, we show how predictive model checks can be used to guide the choice between possible mixture priors. These checks show that extending the mixture model to allow extra variability around zero instead of the usual point mass null fits the data better. Moreover, predictive

email: a.m.lewin@imperial.ac.uk

model checks indicate that such a mixture model is suitable for data pre-processed using the MAS5.0 software, but not using the RMA method, indicating that care has to be taken when applying some of the standard mixture models to RMA processed data.

KEY WORDS: Microarray; Differential Expression; Mixture model; Bayesian analysis; Hierarchical model; MCMC

1. Introduction

This paper deals with a widely-studied problem in microarray analysis, modelling differential gene expression between two experimental conditions. This is usually thought of as a multiple hypothesis testing problem, with one test for each gene, where the null hypothesis for each gene is that it is not differentially expressed. Sometimes the alternative hypothesis is not modelled explicitly, as is the case for various modifications of the t-test (e.g. Baldi & Long, 2001; Tusher et al., 2001) and models using a composite null (Bickel, 2004; Lewin et al., 2006; Bochkina & Richardson, 2007). There have also been approaches where the alternative is modelled explicitly, using mixture models to classify genes as differentially expressed or not. It is this type of model we are concerned with in this work.

A wide variety of mixture models have been proposed, each of which can be expected to give a different classification of the genes in any given data set. Most models are intended to work with data transformed to the log scale, sometimes with shifted log transforms. This is assumed to be the case in the following review of mixture models, except where otherwise indicated.

Mixtures can be specified directly at the data level, e.g. on normalised log fold

changes, or the data can be modelled using a parameter to represent the underlying difference between conditions, which is given a mixture prior. Examples of mixtures at the data level include Broët et al. (2002) who use a mixture of Normals, Parmigiani et al. (2002) and Dean & Raftery (2005) who use a Normal for the null and Uniforms for differentially expressed genes, and Efron et al. (2001), Newton et al. (2001) and Do et al. (2005) who use non-parametric models.

In models formulating a mixture prior on the latent difference parameter, a distributional choice must be made at two levels, the likelihood and the mixture prior. For the likelihood, Lönnstedt & Speed (2003) and Smyth (2004) use a Gaussian, Gottardo et al. (2006) use a t-distribution and Newton et al. (2004) use a Gamma distribution on the original scale of the data. Lönnstedt & Speed (2003) and Smyth (2004) then put a mixture on the log fold change parameter, using a point mass at zero for the null and a Normal (with conjugate prior) for the alternative component. Lönnstedt & Britton (2005) extend this model to use a Normal with small variance for the null. Newton et al. (2004) put a mixture on the pair of parameters representing the gene expression levels in the two conditions. Under the null these parameters are drawn from the same non-parametric distribution, under the alternative they come from different non-parametric distributions. Gottardo et al. (2006) uses the same structure, but with Normal distributions in the mixture.

Out of the above models with mixtures on the parameters, the only one that estimates the parameters of the mixture prior in a fully Bayesian way is the recent work by Gottardo et al. (2006). Other models are fitted using Empirical Bayes methods. The proportion of true null hypotheses is not estimated, thus these

methods produce a ranking of the genes rather than an actual estimate of how many genes are differentially expressed.

In our work we show that it is possible to estimate the proportion of true nulls in a fully Bayesian way. The starting point of our model (described in Section 2) is an ANOVA formulation (Kerr, Martin, & Churchill, 2000), with a Normal likelihood term. As in previous work (Lewin et al., 2006) we stabilise the estimates of gene variances by modelling them as exchangeable within each experimental condition. We use a three-component mixture at the parameter level to model separately the populations of over-expressed, under-expressed and non-differentially expressed genes. It is possible to use a larger number of components to model differential expression, and to estimate the number of components as part of the model (Broët et al., 2002, 2004) but here we focus on the 3-component mixture for ease of interpretation. Individual genes are classified as differentially expressed or not based on their posterior probabilities of being allocated to the various mixture components. Using the posterior probabilities of being differentially expressed we can also estimate the false discovery rate (Benjamini & Hochberg, 1995) for any given list of genes.

In Section 3 we explore the behaviour of a particular form of the model. We model the log fold change parameter with a mixture of a point mass at zero and Gamma distributions for the differentially expressed genes. The reason for this choice is the heavy tail of the Gamma distribution which makes it suitable in the situation where we expect a majority of the log fold changes to be relatively small (in absolute value) but also want to allow for some large values of the log fold change. This formulation is well suited to the experimental context where the gene

expression between different conditions or genotypes is only slightly modified for the majority of changing genes, however the condition investigated may seriously affect a small number of target genes. Also, we do not impose a symmetric prior on the log fold change, since in gene expression experiments, we have no reason to assume that the changes occur equally likely in each direction. Using a simulation study, we show that the model is flexible and behaves well in a number of different situations. In particular, the proportion of true null hypotheses and the false discovery rate can be well estimated.

Despite the flexibility of the model we consider in Section 3, it is undeniable that the choice of the mixture prior can have a large effect on the classification of genes. Simulations cannot tell us the best model to use in general. Even non-parametric approaches are not necessarily "model free", for instance the implicit allocation model used in the Dirichlet Process mixture can be shown to be influential (Green & Richardson, 2001). Hence it is worth developing model checking techniques that can inform the choice of mixture prior for any given data set. In Section 4 we consider a range of alternatives for the mixture prior, in place of the two Gamma distributions for differentially expressed genes we use either two Uniforms (in the spirit of Parmigiani et al., 2002) or one conjugate Normal (as in Lönnstedt & Speed, 2003; Smyth, 2004), and the point mass for the null can be replaced by a Normal with small variance - we call this the nugget null after the terminology introduced in geostatistics to model variograms (Cressie, 1991, p.59). A nugget null was used in Lönnstedt & Britton (2005) in order to reduce the number of differentially expressed genes in a particular data set.

We describe how to use Bayesian mixed predictive p-values to assess the various

combinations of these distributions, for a gene expression data set from wildtype and knock-out mice. Checking the separate components of the mixture distribution is not straightforward, and we have adapted Bayesian p-values to achieve this. For the gene expression data that we analyse, the best model uses the nugget null and the Gamma distributions for the differentially expressed genes. Since other parts of the model may have an effect on the fit of the mixture parameters, we also use predictive checks to look at the specification of the prior for the variances. Details of the alternative models are given in Section 2.

The code used to fit the model is available at http://www.bgx.org.uk.

2. Bayesian Mixture Framework

2.1 *The Model*

We implement a hierarchical model which can be used for analysing paired (2-colour array) or unpaired (single-colour array) differential expression data, or gene expression data under several conditions (multi-class). In this section we give the model for unpaired data; see Section 2.4 for the general likelihood term. For the work in this paper we assume the data has already been normalized. We refer to the model presented in this section as Model 1.

The first level of the model is a Normal distribution for the log gene expression y_{gsr}, for gene g, condition s and replicate array r ($r = 1, ..., R_s$):

$$y_{g1r} \sim N\left(\alpha_g - \delta_g/2, \tau_{g1}\right);\ y_{g2r} \sim N\left(\alpha_g + \delta_g/2, \tau_{g2}\right) \tag{2.1}$$

where α_g is the overall expression level, δ_g is the parameter measuring differential expression and τ_{gs} is precision. Note that we use gene-specific variances $\sigma^2_{gs} \equiv \tau^{-1}_{gs}$.

At the second level, the variances are modelled as exchangeable within each condition, *i.e.* the variances are assumed to come from a common distribution, chosen here to be Inverse Gamma:

$$\tau_{gs} \sim \text{Gam}(a_s, b_s). \tag{2.2}$$

Note that this means that conditionally on a_s, b_s the data y_{gsr} has a t-distribution. We use a flat prior on the overall gene expression levels α_g.

The parameter measuring differential expression, δ_g, has a mixture prior for classification:

$$\delta_g \sim \pi_0 f_0 + \pi_{+1} f_+ + \pi_{-1} f_-, \tag{2.3}$$

where f_0 is the density of the null component and f_- and f_+ are the densities for genes which are under-expressed and over-expressed respectively. The parameters $\pi_0, \pi_{-1}, \pi_{+1}$ represent the proportions of genes in the three components, and sum to 1.

We consider a number of different parametric families for the mixture components. Our main focus is on a model with a point mass at zero for the null component, a Gamma distribution for over-expressed genes and a 'reverse' Gamma for the under-expressed genes:

$$\delta_g \sim \pi_0 \delta_0 + \pi_{+1} \text{Gam}^{(+)}(\lambda_+, \eta_+) + \pi_{-1} \text{Gam}^{(-)}(\lambda_-, \eta_-), \tag{2.4}$$

where $\text{Gam}^{(+)}$ indicates the usual Gamma density function and the density $\text{Gam}^{(-)}$

is defined on the negative real line ($x \sim \text{Gam}^{(-)}$ means $-x \sim \text{Gam}^{(+)}$). Fixing the shape parameter of the Gamma distributions to be greater than 1 helps to separate the posterior distributions of δ_g under the null and alternative hypotheses, thus leading to a clearer classification. Here we use $\lambda_+ = \lambda_- = 1.5$, unless otherwise specified.

The third level of the model consists of the priors on the hyperparameters. The variance hyperparameters a_s, b_s are given Gamma(0.01,0.01) priors. With the large amount of data used here we have found that results on τ_{gs} are not sensitive to this choice. For a more in-depth discussion of the choice of prior for b_s see the supplementary material of Bochkina & Richardson (2007). For the hyperparameters η_+, η_- we use the more informative Exp(1) distribution. This ensures that the mixture can be fitted even when there are no differentially expressed genes. The mixture weights $\pi_0, \pi_{-1}, \pi_{+1}$ are given a Dirichlet distribution with parameters $1, 1, 1$. This model is explored in a simulation study in Section 3.

In the model checking Section 4 we consider two alternative models for differentially expressed genes: one using Uniform distributions $f_+ = U(0, \eta_+)$ and $f_- = U(\eta_-, 0)$, η_- and η_+ fixed, and the "conjugate" model with just one two-sided component $f_+ = N(0, c\tau_g)$, with c a parameter of the model, and τ_g the precision given in 2.1 with the constraint $\tau_{gs} = \tau_g$. We also consider an alternative for the null component: $f_0 = N(0, \tau_\epsilon)$ where τ_ϵ is fixed. We use the model-checking procedure to determine the best value for τ_ϵ. In order to test the specification of the hierarchical model for the variance (2.2), we perform checks on a log Normal prior for the τ_{gs} as well as the Gamma prior. Table 1 shows all the alternative models we consider.

Alternative mixture priors (DE genes)	Hyper-priors
$f_+ = \text{Gam}^{(+)}(\lambda_+, \eta_+), f_- = \text{Gam}^{(-)}(\lambda_-, \eta_-)$	$\lambda_\pm$ fixed, $\eta_\pm \sim Exp(1)$
$f_+ = \text{Unif}(0, \eta), f_- = \text{Unif}(-\eta, 0)$	η fixed
$f_+ = N(0, c\tau_g)$	$c \sim \text{Gam}(1, 0.5)$
Alternative mixture priors (null genes)	Hyper-priors
$f_0 = \delta(0)$	-
$f_0 = N(0, \tau_\epsilon)$	τ_ϵ fixed
Alternative variance priors	Hyper-priors
$\tau_{gs} \sim \text{Gam}(a_s, b_s)$	$a_s, b_s \sim Gam(0.01, 0.01)$
$\tau_{gs} \sim \text{logNormal}(a_s, b_s)$	$a_s \sim N(0, 0.001), b_s \sim Gam(0.01, 0.01)$

Table 1

Alternative Models considered in Section 4 (The model in the first line is Model 1). Parameters which are given as fixed, are chosen based on the results of the model checks.

2.2 *Rules for Selecting Genes*

In order to estimate the mixture model, it is usual to write the model using latent allocation variables. These are defined as $z_g = 0, 1, -1$ when gene g is allocated to the mixture component for non-DE genes, over- and under-expressed genes respectively. The prior for z_g is $\mathbb{P}(z_g = i) = \pi_i$ for $i = 0, 1, -1$.

These allocation parameters are used to classify the genes into different mixture components. The posterior probability of being in the null component, for example, is $p_g \equiv \mathbb{P}(z_g = 0|\text{ data })$. A natural decision rule for classification is as follows: a gene is declared DE if $p_g < p_{cut}$, where the cut-off on the posterior probabilities must be chosen according to some loss function. The choice $p_{cut} = 0.5$ is the Bayes Rule. This corresponds to equal penalties for false positives and false negatives. Other values for the cut-off correspond to more or less conservative choices. Note that for a mixture with more than two components such as the one

we are discussing, more complicated rules may be appropriate if there is a lot of variability in the data (see Discussion for details).

2.3 *Estimating False Discovery Rate*

A simple estimate of the false discovery rate (FDR) can be obtained using the posterior probabilities p_g (Newton et al., 2004; Broët et al., 2004):

$$\widehat{FDR} = \frac{1}{N_{p_{cut}}} \sum_{g \in S_{p_{cut}}} p_g, \tag{2.5}$$

where $S_{p_{cut}}$ is the set of genes declared DE, i.e. such that $p_g < p_{cut}$, and $N_{p_{cut}}$ is the number of genes in $S_{p_{cut}}$. The false non-discovery rate (FNR), which is the expected number of false negatives among those declared non-DE, can be estimated in a similar way, by summing over $1 - p_g$ for the genes declared non-DE.

2.4 *Implementation*

We estimate the model in a fully Bayesian way, using our own code written in C++ to perform Monte Carlo Markov Chain (MCMC) simulations of the posterior distribution (available at http://www.bgx.org.uk). The program implements a more general parametrization of the model than that given in (2.1). In this version, the first level of the model is expressed as

$$y_{gsr} \sim N\left(\sum_l \beta_{gl} x_{ls}, \tau_{gs}\right). \tag{2.6}$$

The index s is equal to one in the paired differential expression case and labels experimental condition for unpaired differential expression and multi-class data. The

x_{ls} are fixed constants which determine the parametrization of the gene means. The stochastic parameters are the β_{gl}. In the paired case $l = 1$, $x_{11} = 1$ and $\beta_{g1} \equiv \delta_g$, the parameter which represents log fold change. For unpaired differential expression data $s = 1, 2$ and $l = 1, 2$, with the matrix $x = \begin{pmatrix} 1 & 1 \\ -1/2 & 1/2 \end{pmatrix}$. In this case we have a gene effect or overall expression level $\beta_{g1} \equiv \alpha_g$ and a log fold change parameter $\beta_{g2} \equiv \delta_g$. For more than 2 experimental conditions, it is convenient to use the identity matrix for x, and the expression level in condition s is then β_{gs}.

At most one of the β_{gl} can have a mixture prior, the rest have flat priors. The program may be run with flat priors on all the β_{gl}. The update for the allocation parameters z_g is a reversible jump type move: we must update δ_g at the same time as z_g since effectively the dimension of the space for δ_g changes with z_g. Gottardo & Raftery (2004) have proposed a similar type of move for a mixture involving a point mass and a continuous distribution. Details of our update for δ_g and z_g are given in the Supplementary material.

3. Simulation Study: Estimation of the Proportion of Differentially Expressed Genes

The model is intended to be flexible and able to fit many different data sets. Here we investigate the performance of the model for a variety of patterns of differential expression. We have simulated data for 2500 genes, under 2 experimental conditions with 8 replicate arrays for each condition (unless stated otherwise below). The data comes from a Normal distribution, as in (2.1). The variances σ^2_{gs} are log Normal: $\log(\sigma^2_{gs}) \sim N(-1.8, 0.5)$, with the values of the parameters coming

from a least squares fit to real data. The gene means α_g come from a wide Normal distribution on the same scale as real data (mean 7 and standard deviation 5). We vary the distribution for the δ_g of the differentially expressed genes, as detailed in the following subsections. The basic distribution is as follows:

$$\begin{aligned} \delta_g \sim \pi_0\delta_0 \quad &+ \quad \pi_{+1}(\phi\text{Unif}(0.07, 0.7) + (1-\phi)N(0.7, 0.8)) \\ &+ \quad \pi_{-1}(\phi\text{Unif}(-0.7, -0.07) + (1-\phi)N(-0.7, 0.8)), \end{aligned} \tag{3.1}$$

except for in the Asymmetric Case below. In the following subsections, the model was fitted for 50 simulated data sets in each set-up. The 50 simulations use the same α_g, δ_g and σ^2_{gs} but simulate the data y_{gsr} 50 times.

3.1 *Distribution of Differentially Expressed Genes*

The use of the mixture of Uniform and Normal distributions for simulating the DE genes is to ensure we are not simulating from the model we fit and to test how the Gamma mixture prior works when the data is far from a standard peaked distribution. We have simulated 3 Cases with $\phi = 0.3, 0.5$ and 0.8, Cases 1, 2 and 3a respectively, to give increasing non-Normality of DE genes. Here the true value of π_0 is 0.8.

The upper row in Figure 1 shows the true δ_g and σ_g parameters in the three cases, showing the increasing uniformity of the alternative distributions. The curves superimposed are the alternative densities estimated in the model, using the posterior mean (averaged over the 50 simulations) of the hyperparameters η_+ and η_-. (The y-axes for the curves are arbitrary.) The bottom row shows one realisation of each corresponding data standard deviation and difference between

conditions.

The model fits well to the data, finding average posterior means for π_0 of 0.805, 0.797 and 0.781 respectively for the 3 cases of increasing ϕ. Table 2 shows the credibility ranges containing 95% of the posterior means amongst the 50 simulations. For the most non-standard distribution (Case 3a), the estimates of π_0 are less good, but still close to the true value of 0.8. The table also shows the posterior standard deviations of π_0. Conditional on the model being an adequate representation of the data, we can estimate the proportion of true nulls with high precision.

3.2 *Number of Differentially Expressed Genes*

The model also does very well when there are fewer DE genes. We have looked at the case where ϕ is 0.8 (the most non-standard alternative distribution for the DE genes) and varied π_0. We use 4 Cases here: true value of π_0 is 0.8, 0.95, 0.99 and 1. These are Cases 3a-3d. The estimates of π_0 are shown in Table 2. Even when π_0 is close to 1, it is estimated well. On average, π_0 is estimated to be 0.999 when the true value is 1. In this "null" case, using Bayes Rule, the average number of differentially expressed genes over the 50 simulations is 0.12.

The estimate of π_0 gets better with increasing π_0, since the form of the prior on the δ_g has less effect and thus the estimate of π_0 is better. When π_0 is smaller there is more information to fit this distribution, hence any mis-specification is more influential.

	True π_0	Posterior mean π_0	Posterior SD π_0	Comment
Case 1	0.8	0.805 (0.798, 0.811)	0.010 (0.009, 0.010)	most peaked alternative
Case 2	0.8	0.797 (0.792, 0.802)	0.010 (0.009, 0.010)	less peaked alternative
Case 3a	0.8	0.781 (0.771, 0.789)	0.011 (0.011, 0.012)	least peaked alternative
Case 3b	0.95	0.947 (0.941, 0.951)	0.007 (0.006, 0.007)	least peaked alternative
Case 3c	0.99	0.990 (0.988, 0.992)	0.003 (0.002, 0.003)	least peaked alternative
Case 3d	1	0.999 (0.999, 0.999)	0.001 (0.001, 0.001)	no DE genes
Case 3b*	0.95	0.956 (0.945, 0.967)	0.010 (0.008, 0.012)	only 3 replicates
Case 3c*	0.99	0.996 (0.993, 0.998)	0.003 (0.002, 0.004)	only 3 replicates
Case 4	0.9	0.897 (0.893, 0.900)	0.007 (0.006, 0.007)	asymmetric alternatives

Table 2

Simulation results. Column three gives the posterior mean π_0 averaged over the 50 simulations, along with the interval containing 95% of the posterior means (in brackets). Column four gives the same for the posterior standard deviations (averaged on the variance scale).

3.3 *Number of Replicate Arrays*

We have repeated Cases 3b and 3c using only 3 replicates in each experimental condition (Cases 3b* and 3c*). Results for π_0 are also shown in Table 2. As expected, there is more variability in the estimates, but the posterior means of π_0 are still close to the true values.

3.4 *Asymmetry of Over and Under Expressed Genes*

The last situation (Case 4) we consider is one where over- and under-expressed genes have different distributions. For this we simulated the differential effects as

$$\begin{aligned} \delta_g \sim \pi_0\delta_0 \quad &+ \quad \pi_{+1}(\phi\text{Unif}(0.01, 1.7) + (1-\phi)N(1.7, 0.8)) \\ &+ \quad \pi_{-1}(\phi\text{Unif}(-0.7, -0.01) + (1-\phi)N(-0.7, 0.8)) \end{aligned} \tag{3.2}$$

and use $\phi = 0.6$. The values for π_0, π_{+1} and π_{-1} are 0.9, 0.09 and 0.01 respectively.

The average estimates for π_0, π_{+1} and π_{-1} are 0.897, 0.093 and 0.011. Figure 2 shows the true δ_g and σ_g parameters, with the Gamma curves superimposed, as before. As expected, the separate Gamma distributions model an asymmetric distribution of differentially expressed genes well.

3.5 *Estimating the False Discovery Rate*

In the simulations above, we can calculate the true error rates since we know which genes are truly differentially expressed. Therefore we can assess our estimates of the FDR and FNR. Recall that we declare gene g to be DE if $p_g < p_{cut}$. In order to compare with the true rates, we calculate FDR and FNR over a range of p_{cut}. Figure 3 shows true and estimated error rates for Case 1 and Case 3a, which are the cases with the best and worst estimated rates. The FDR is estimated in both cases very well, while the FNR is slightly over-estimated in Case 3a.

4. Model Checking

Though we have shown that our 3-component Gamma model can obtain good estimates of π_0 in a range of situations, it is nevertheless clear that the choice of parametric family for the mixture can affect the classification of genes. In this section we present results on a data set consisting of wildtype and knock-out mice, showing strikingly different classifications when using different mixture models. We then show how Bayesian predictive checks can give some useful guidance for choosing between different models.

4.1 *Data Sets*

The data set consists of gene expression levels in 6 knock-out and 5 wildtype mice, measured using Affymetrix microarrays. The gene knocked out is the IRS2 gene. For more details of this data set, see Craig et al. (submitted). We have pre-processed the data using both the MAS5.0 software (Hubbell et al., 2002) and the RMA package (Irizarry et al., 2003), in both cases using loess non-linear normalization between arrays. In Section 4.5 we show that the class of hierarchical mixture models presented here is not suitable for the RMA-processed data, since RMA gives small variance estimates.

4.2 *Predictive Model Checks*

In order to perform model checks, we use Bayesian predictive p-values which "summarise" the level of agreement between the data and predictions from the model. To be precise, we compute Bayesian p-values through a mixed predictive distribution (Gelman et al., 1996; Marshall & Spiegelhalter, 2003), which is a posterior predictive distribution conditional on the global hyperparameters of a hierarchical model rather than conditional on the estimated gene-specific likelihood parameters. The predictive process can be thought of as predicting new gene-specific parameters from their prior distribution followed by new data points, hence the name "mixed" predictive.

The usual posterior predictive p-values based on the likelihood are known to be very conservative due to the double use of the data (Bayarri & Berger, 2000). Mixed predictive p-values (whilst still posterior predictive quantities) are much less conservative since the influence of each data point on the hyperparameters is much weaker than its influence on the likelihood parameters (Marshall & Spiegel-

halter, 2003).

Different aspects of the model can be checked with different checking functions. The quantities we choose to predict are the sums of squares for each gene and condition $S_{gs} \equiv \frac{1}{(R_s-1)} \sum_r (y_{gsr} - \bar{y}_{gs.})^2$, which correspond to the gene variances σ^2_{gs} (as in Lewin et al., 2006), and the mean differences between conditions for each gene $d_g \equiv \bar{y}_{g2.} - \bar{y}_{g1.}$, which correspond to δ_g.

4.3 *Checking Gene Variances*

The sums of squares are straightforward to predict. In the hierarchy of the model, S_{gs} is dependent on just one set of gene-specific parameters, τ_{gs}, and these are dependent on the hyperparameters a_s, b_s. Therefore the mixed predictive density of $S^{mixpred}_{gs}$ is

$$f(S^{mixpred}_{gs}|\mathbf{y}^{obs}) = \int f(S^{mixpred}_{gs}|a_s, b_s)\pi(a_s, b_s|\mathbf{y}^{obs})da_s db_s, \tag{4.1}$$

where $\pi(a_s, b_s|\mathbf{y}^{obs})$ is the posterior density of a_s, b_s and $f(S^{mixpred}_{gs}|a_s, b_s)$ can be found by predicting new parameters τ^{pred}_{gs}:

$$\begin{aligned} \tau^{pred}_{gs} &\sim \text{Gam}(a_s, b_s), \\ S^{mixpred}_{gs}|\tau^{pred}_{gs} &\sim Gam\left(\frac{1}{2}(R_s-1), \frac{1}{2}(R_s-1)\tau^{pred}_{gs}\right). \end{aligned} \tag{4.2}$$

In practice this distribution is simulated in the MCMC run at the same time as fitting the model. This is done by predicting a new τ^{pred}_{gs} at each MCMC iteration conditional on the current values of a_s, b_s, and predicting $S^{mixpred}_{gs}$ conditional on τ^{pred}_{gs} (as shown in (4.2)).

The Bayesian p-values are then defined as

$$\mathbb{P}(S_{gs}^{mixpred} > S_{gs}^{obs} | \mathbf{y}^{obs}), \tag{4.3}$$

where S_{gs}^{obs} is the observed sum of squares. These quantities should be close to Uniform for each s, if the model has a reasonable fit to the data.

For the data described above (processed using MAS5.0), Figure 4 (a) shows the predictive p-values for the first of the two conditions (plots for both conditions look similar). For most of the range the p-values are flat, but with a small excess of p-values towards low values. We have not been able to find a parametric model for the gene-specific variability giving better results than those shown for this data set. A log-Normal distribution (as shown in Table 1) gives very similar results, with slightly more variability (Figure 4 (b)). In Figure 4 (c), the mixed predictive p-values for the RMA-processed data with inverse gamma prior is given. These are discussed in Section 4.5.

4.4 *Checking Mixture Prior*

For the mixture part of the model, implementing a mixed predictive approach is not so straightforward. The observed differences d_g are dependent on three sets of gene-specific parameters: δ_g, τ_{gs} and z_g. Should we integrate over all three sets in order to condition on the global hyperparameters?

Clearly we need to integrate over δ_g, as this is the parameter which most affects d_g, therefore we predict new δ_g^{pred}. Re-predicting the τ_{gs} has little effect on the Bayesian p-values for d_g, so we condition on these in all results shown here. We do not integrate over the mixture allocations z_g either. Since we wish to interpret the

mixture components separately (i.e. when we use them to classify the genes), we should check the components separately. Thus we condition on the z_g, and do not re-predict these parameters when calculating the mixed predictive distribution.

Hence the prediction scheme for the mixture part of the model is

$$\begin{aligned} \delta_g^{pred} &\sim \pi_0\delta_0 + \pi_{+1}\mathrm{Gam}^{(+)}(\lambda_+,\eta_+) + \pi_{-1}\mathrm{Gam}^{(-)}(\lambda_-,\eta_-), \\ \bar{y}_{gs.}^{mixpred} &\sim N\left(\alpha_g \mp \delta_g^{pred}/2, R_s\tau_{gs}\right), \\ d_g^{mixpred} &= \bar{y}_{g2.}^{mixpred} - \bar{y}_{g1.}^{mixpred}, \end{aligned} \tag{4.4}$$

Note that in the unpaired case, the fixed effect α_g is not re-predicted.

The Bayesian p-values are then defined as $\nu_g \equiv \mathbb{P}(d_g^{mixpred} > d_g^{obs}|\mathbf{y}^{obs})$. However, for checking the mixture components, it is not straightforward how to use these quantities. If we consider each gene to have come from one of the mixture components, the ν_g can be thought of as a mixture of p-values from different distributions, one Uniform for the correct classification and two non-Uniform when the gene is mis-classified, so we would not expect to see Uniform ν_g even for the true model. To overcome this problem, we need to separate out the three mixture components. There are two stages to this process. First we condition each p-value on the allocation of the gene:

$$\nu_{g,j} \equiv \mathbb{P}(d_g^{mixpred} > d_g^{obs}|\mathbf{y}^{obs}, z_g = j) \tag{4.5}$$

for $j = 0, 1, -1$. This separates out the contributions of the different mixture components to each Bayesian p-value. Second, when we plot histograms of $\nu_{g,j}$

(separately for each j) we restrict the plots to genes with high posterior probability of being in component j, i.e. $\mathbb{P}(z_g = j|\mathbf{y}^{obs}) > \rho_j$. This second step reduces the contribution of the mis-classified genes. We have investigated these quantities by simulation in a separate paper (Lewin and Richardson, in preparation), including dependence on the choice of ρ_j. We expect a deficit of extreme p-values due to the overlap of mixture components, but the shapes of the distributions give a good indication of the fit of the model.

Here we show the plots of $\nu_{g,j}$ for three different mixture models, to illustrate how we can gain insight into the fit of the mixture using these quantities. First we fit a mixture model with Uniform distributions in place of the Gamma distributions for the over and under-expressed genes:

$$\delta_g \sim \pi_0\delta_0 + \pi_{+1}\text{Unif}(0,\eta) + \pi_{-1}\text{Unif}(-\eta,0). \tag{4.6}$$

The results shown here are for $\eta = 3$, but we get similar shapes for the predictive p-values using other values. Figure 5 shows $\nu_{g,j}$ in this case (only for genes with $\mathbb{P}(z_g = j|\mathbf{y}^{obs}) > 0.5$). The plot of $\nu_{g,0}$ shows an excess of extreme values, suggesting that the null component is too narrow for this data. The outer components have very skewed distributions. Note that the shape of the distribution indicates the shape of the difference between observed (d_g^{obs}) and predicted (d_g^{pred}) data. The small p-values for the left-hand component correspond to genes with δ_g near zero, thus the skewness of the $\nu_{g,-1}$ to the left indicates a skewness of δ_g towards the centre, compared with δ_g^{pred}. This suggests that we should replace the Uniform with a distribution on the δ_g giving less weight to the extremes (away from zero).

Hence, we next fit the main model given in Section 2, with Gamma mixture components given in (2.4). The results shown here use $\lambda_+ = \lambda_- = 5$, but other values give similar shapes for the predictive p-values distributions. Figure 6 shows histograms of $\nu_{g,j}$. Now the distributions of $\nu_{g,-1}$ and $\nu_{g,1}$ are skewed in the opposite direction from before. The deficit of small $\nu_{g,-1}$ and large $\nu_{g,1}$ might be partly due to the overlap between mixture components and the fact we are only plotting these quantities for genes with $\mathbb{P}(z_g = j|\mathbf{y}^{obs}) > 0.5$. However, the shapes indicate systematic departure from uniformity. Further, the null component still has some excess of extreme p-values.

To get more Uniform $\nu_{g,j}$ for the null mixture component, we thus expand our model and replace the point mass at zero in (2.4) with a nugget null:

$$\delta_g \sim \pi_0 N(0, \tau_\epsilon) + \pi_{+1}\text{Gam}^{(+)}(\lambda_+, \eta_+) + \pi_{-1}\text{Gam}^{(-)}(\lambda_-, \eta_-), \tag{4.7}$$

where the precision parameter τ_ϵ is fixed to a reasonably large value (corresponding to a small variance). We have used several different values of τ_ϵ, and used the plot of predictive p-values to choose the best value for τ_ϵ. The results shown in Figure 7 are for $\tau_\epsilon = 100$. These results show a much better fit to the data than for the previous models, with a close to Uniform histogram of $\nu_{g,0}$. There are fewer genes classified as DE, so the histograms for the outer components are difficult to judge, but the q-q plots show a good agreement with Uniformity for the left and central components (see Table 1).

We have also tried the conjugate Normal mixture prior for this data set. As the non-null mixture component extends on both sides of the null, we use 2-sided p-

values. With the point null, this model again exhibits too many extreme p-values for the null component. Using the nugget null improves the prediction for that component, but the outer components do not fit as well with the one-component Normal as with the two Gamma components. Note that the latter model (nugget null and normal alternative) is essentially the model used by Lönnstedt & Britton (2005).

The models considered here give very different results for classification and probability of the null: the Uniform mixture obtains $\hat{\pi}_0 = 0.96$, the Gamma model with the point null obtains $\hat{\pi}_0 = 0.68$, the Gamma with the nugget null has $\hat{\pi}_0 = 0.99$, the conjugate model with delta-function prior has $\hat{\pi}_0 = 0.51$ and the conjugate model with the nugget null obtains $\hat{\pi}_0 = 0.94$. The Gamma + nugget null model chosen by the predictive model checks gives a sensible estimate for π_0, with 61 out of 22685 genes declared differentially expressed when the Bayes Rule is used for classification.

4.5 *Different choices for pre-processing the data*

We have attempted to find a hierarchical formulation which fits the RMA processed data well, but we have not been able to do so. Figure 8 shows the $\nu_{g,j}$ for the model specified by (2.1), (2.2) and (2.4). There is a large excess of values close to 0.5 for the null, which indicates over-fitting of the model. We have tried to rectify this by using the nugget null as in the previous section, but this makes virtually no visible difference to the histogram of p-values. The predictive p-values for the sums of squares are also more shrunk towards the centre of the range than for the MAS5.0 data (see Figure 4 (c)).

The lack of fit of the null component here probably indicates we need a dis-

tribution for the data which is narrower than the Normal-Gamma distribution (we are in the opposite situation from the usual case of needing to accommodate outlying data points). This makes sense as the RMA method is known to give smaller variance estimates than the MAS5.0 method.

5. Discussion

We have developed a mixture model framework for classifying genes as differentially expressed between different experimental conditions, using a range of parametric distributions for modelling the different groups of genes. Parametric models with predictive model checks are an alternative to non-parametric models. Minor irregularities in the data can be smoothed out. But it is essential to get some idea of whether the model is a reasonable fit.

For this purpose, we have found that amongst the mixture models investigated, a mixture with a nugget null fits the data better than a point mass at zero. This may be related to the results of Efron (2004) who estimated the empirical null in a mixture model on transformed p-values. He found that the theoretical null, a standard Normal, did not fit the data well. The empirical null was a Normal with non-zero mean and variance slightly larger than one. The nugget null has also been explored in a theoretical context by Rousseau (2006).

For data processed with the RMA method, we have not been able to find any mixture prior which works with the Normal-Gamma distribution for the errors. This indicates that it may not be appropriate to use the Normal-Gamma distribution for data processed by this method.

The results for the predictive model checks are conditional on the use of the

Normal-Gamma distribution at the data level. However, the classification of genes depends on the combination of the distributions at the first and second levels. We have found a particular combination of distributions fit the data well, but there may be other combinations which would also give a good fit.

The model criticism approach that we have developed is suited to checking the form of the distributions used in the different hierarchical models. Other aspects of a model can also affect the classification of genes. In cases where we cannot tell which model gives a better fit using the predictive checks, we could also use the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) .

In Section 2.2 we suggested using a classification rule which declares genes to be differentially expressed if the posterior probability $p_g \equiv \mathbb{P}(z_g = 0|$ data $)$ is less than some threshold value. When there are more than two mixture components this rule may not always be appropriate. For example, consider a gene with almost equal posterior probabilities for being allocated into each of the three mixture components. The Bayes Rule would declare this as differentially expressed, but it could be argued that this gene should be classified as non-differentially expressed, as it has equal evidence for under and over-expression. To deal with this situation we could propose a rule which declares genes to be differentially expressed only if the posterior probability of being classified into *one* of the non-null components is greater than some threshold, e.g. 0.5 or 0.8. Thus genes with equal probabilities of allocation to all mixture components will be declared as non-differentially expressed.

Acknowledgements

The authors would like to thank Marta Blangiardo, Peter Green and Anne-Mette Hein for statistical discussions, Dominic Withers for providing the IRS2 data set and the BAIR group (BAIR project, www.bair.org.uk) for sharing their biological insights. This work was completed while AL and SR were supported by the BBSRC "Exploiting Genomics" grant 28EGM16093 and NB was funded by a Wellcome Trust Cardio - Vascular grant 066780/Z/01/Z. The authors gratefully acknowledge the Wellcome Trust Functional Genomics Development Initiative (FGDI) thematic award "Biological Atlas of Insulin Resistance (BAIR)", PC2910 DHCT, which has supported the generation of the data used in this paper.

References

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

Bayarri, M. J. and Berger, J. (2000). P-values for composite null models. *Journal of the American Statistical Association* **95**, 1127–1142.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Bickel, D. R. (2004). Degrees of differential gene expression: detecting biologically significant expression differences and estimating their magnitudes. *Bioinformatics* **20**, 682–688.

Bochkina, N. and Richardson, S. (2007). Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics* in press.

Broët, P., Lewin, A., Richardson, S., Dalmasso, C., and Magdelenat, H. (2004). A mixture model based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics* **20(16)**, 2562–2571.

Broët, P., Richardson, S., and Radvanyi, F. (2002). Bayesian Hierarchical Model for Identifying Changes in Gene Expression from Microarray Experiments. *Journal of Computational Biology* **9**, 671–683.

Cressie, N. 1991, Statistics for Spatial Data (John Wiley).

Dean, N. and Raftery, A. E. (2005). Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC Bioinformatics* **6**.

Do, K.-A., Mueller, P., and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Applied Statistics* **54**, 627–644.

Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Society* **99**, 96–104.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica* **6**, 733–807.

Gottardo, R. and Raftery, A. E. (2004). Markov chain Monte Carlo computations with mixture of singular distributions. Technical Report 470, Statistics Department, University of Washington, Seattle.

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. E. (2006). Bayesian Robust Inference for Differential Gene Expression in Microarrays with Multiple Samples. *Biometrics* **62**, 10–18.

Green, P. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.

Hubbell, E., Liu, W., and Rui, M. (2002). Robust estimators for expression analysis. *Bioinformatics* **18**, 1585–1592.

Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B., , and Speed, T. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**, e15.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **8**, 819–837.

Lewin, A., Richardson, S., Marshall, C., Glazier, A., and Aitman, T. (2006). Bayesian Modelling of Differential Gene Expression. *Biometrics* **62**, 1–9.

Lönnstedt, I. and Britton, T. (2005). Hierarchical Bayes models for cDNA microarray gene expression. *Biostatistics* **6**, 279–291.

Lönnstedt, I. and Speed, T. (2003). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine* **22**, 1649–1660.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2001). On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology* **8**, 37–52.

Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics* **5**, 155–176.

Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society B* **64**, 1–20.

Rousseau, J. (2006). Approximating Interval hypothesis : p-values and Bayes factors. Technical Report, Université Paris Dauphine.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, 3.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 1–34.

Tusher, V., Tibshirani, R., and Gilbert, C. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences, USA* **98**, 5116–5121.

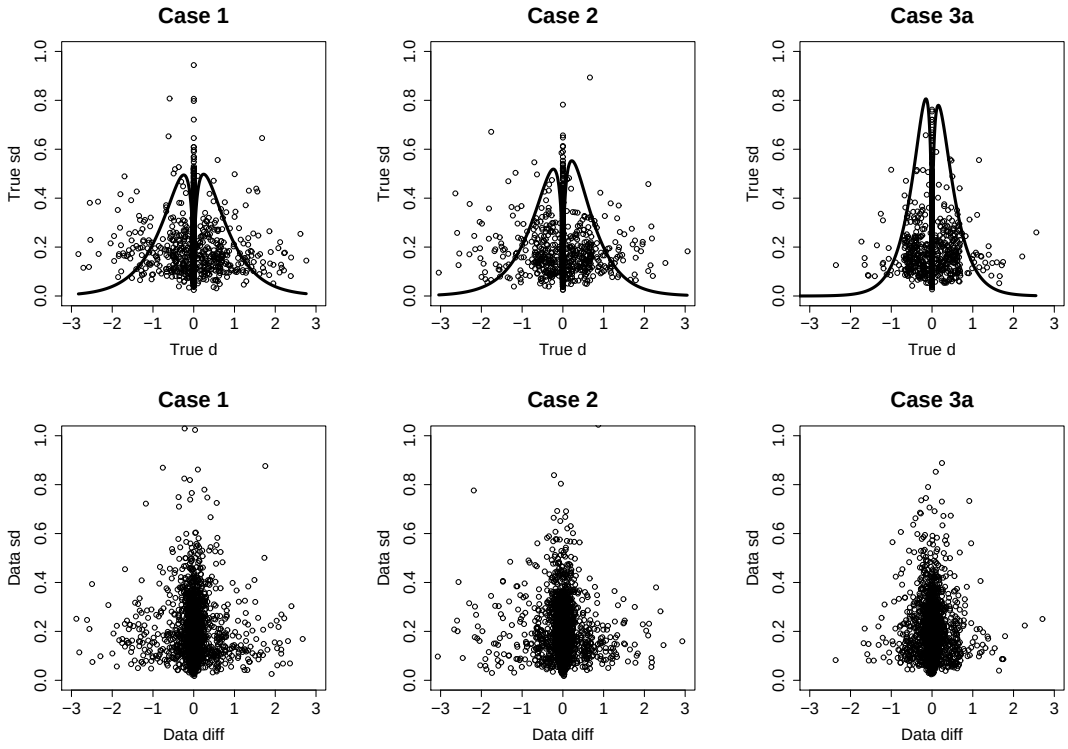

Figure 1. Upper row: true σ_g versus true δ_g for Cases 1, 2 and 3a. Superimposed curves are the estimated Gamma densities for the alternative distributions (using posterior means of η_+, η_-). Bottom row: one realisation of the data standard deviations and difference between conditions.

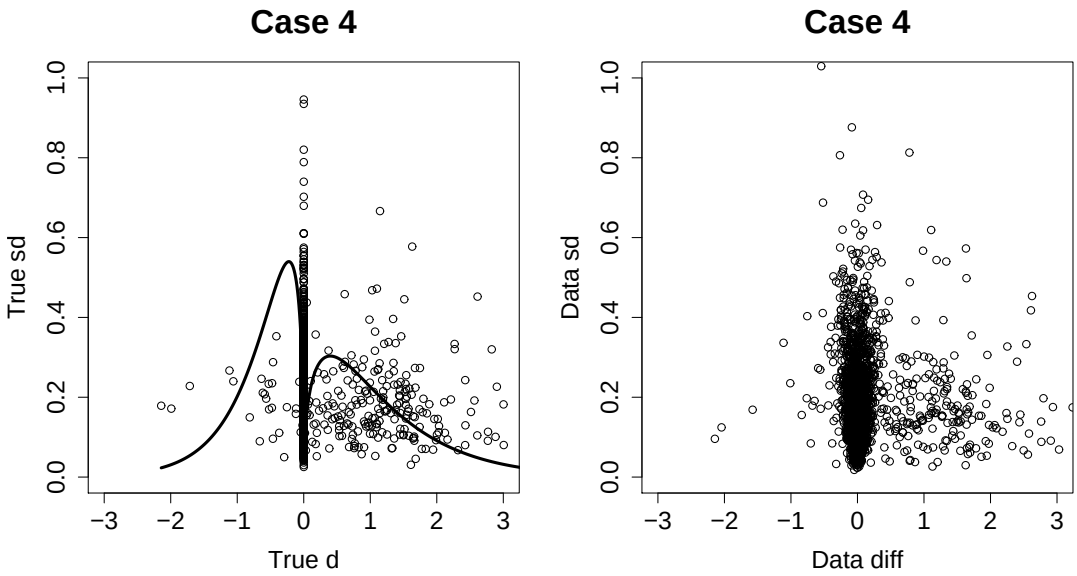

Figure 2. Left: true σ_g versus true δ_g for Case 4. Superimposed curves are the estimated Gamma densities for the alternative distributions (using posterior means of η_+, η_-). Right: one realisation of the data standard deviations and difference between conditions.

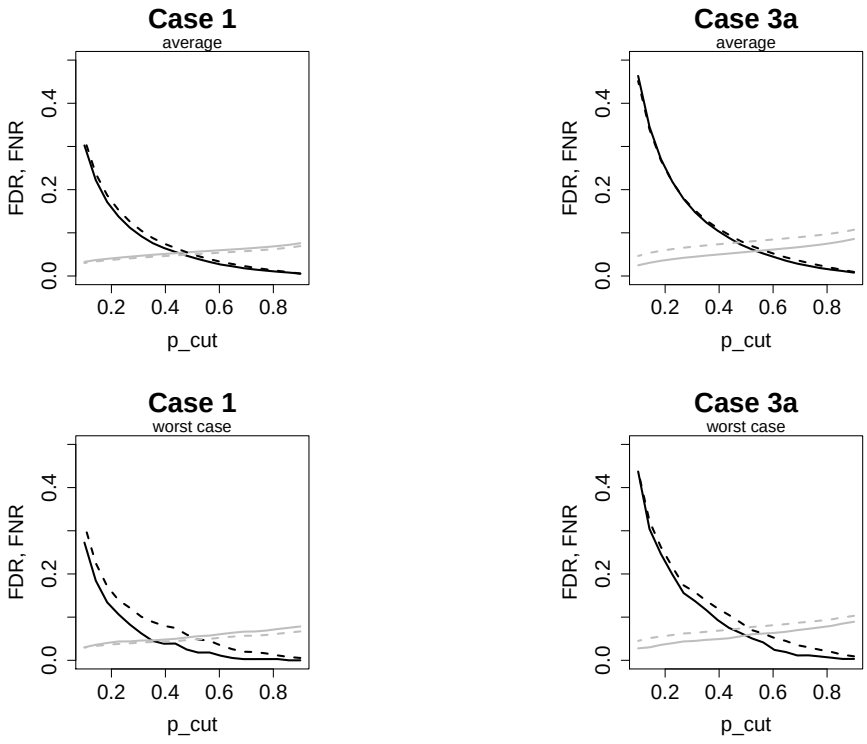

Figure 3. FDR (black) and FNR (grey) for simulation Cases 1 and 3a. Solid lines are true values, dashed lines are estimated. The top two plots show curves averaged over 50 simulations. The bottom plots show the curves for the simulation with the largest discrepancy between true and estimated error rates.

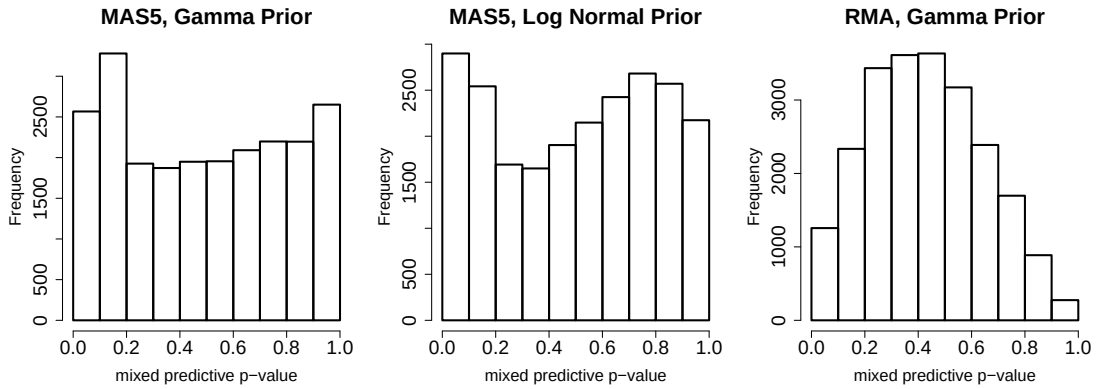

Figure 4. Mixed predictive p-values for sums of squares (wildtype mice, IRS2 data). (a) Gamma prior on variances, MAS5-processed data, (b) Log Normal prior on variances, MAS5-processed data, (c) Gamma prior on variances, RMA-processed data

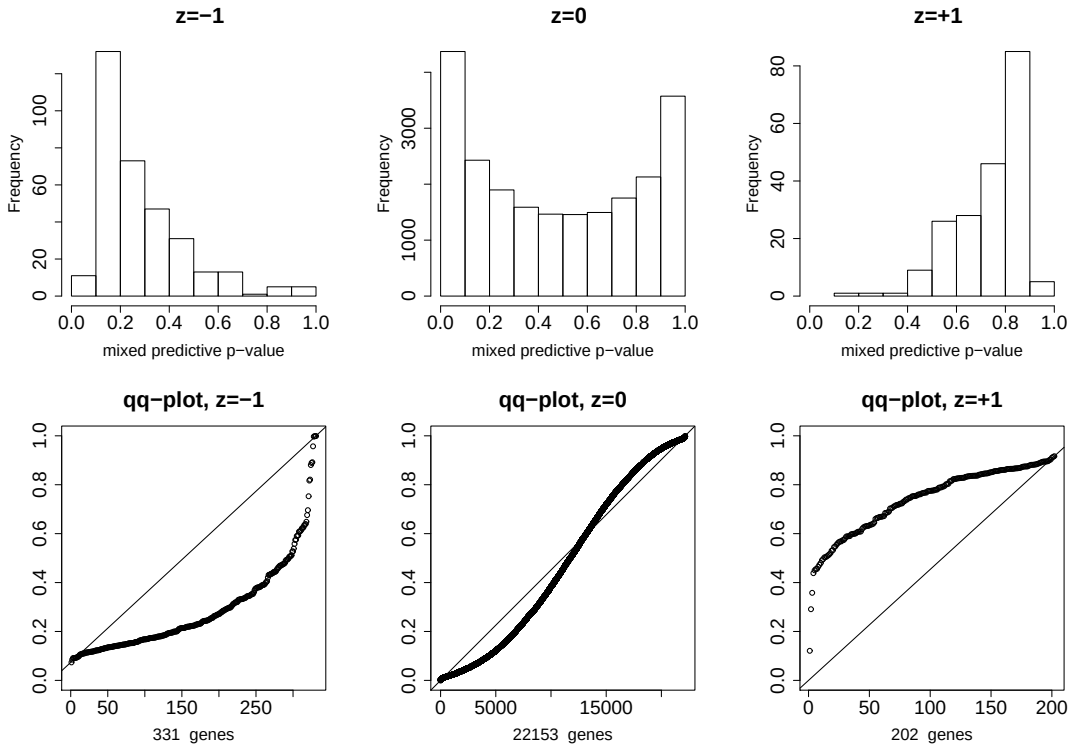

Figure 5. Mixed predictive checks for the mixture of delta-function and Uniforms. Mixed predictive p-values for mean differences conditional on z_g (i.e. $\nu_{g,j}$) for $j = -1, 0, 1$. First row shows histograms, second row shows quantile-quantile plots of the distributions versus a Uniform distribution. The histogram and q-q plot for component j is made from genes with $\mathbb{P}(z_g = j|\mathbf{y}^{obs}) > 0.5$. The number of such genes is shown below the q-q plots. Data shown is the wildtype mice, IRS2 data, fitted with the model using the mixture of delta-function and Uniforms.

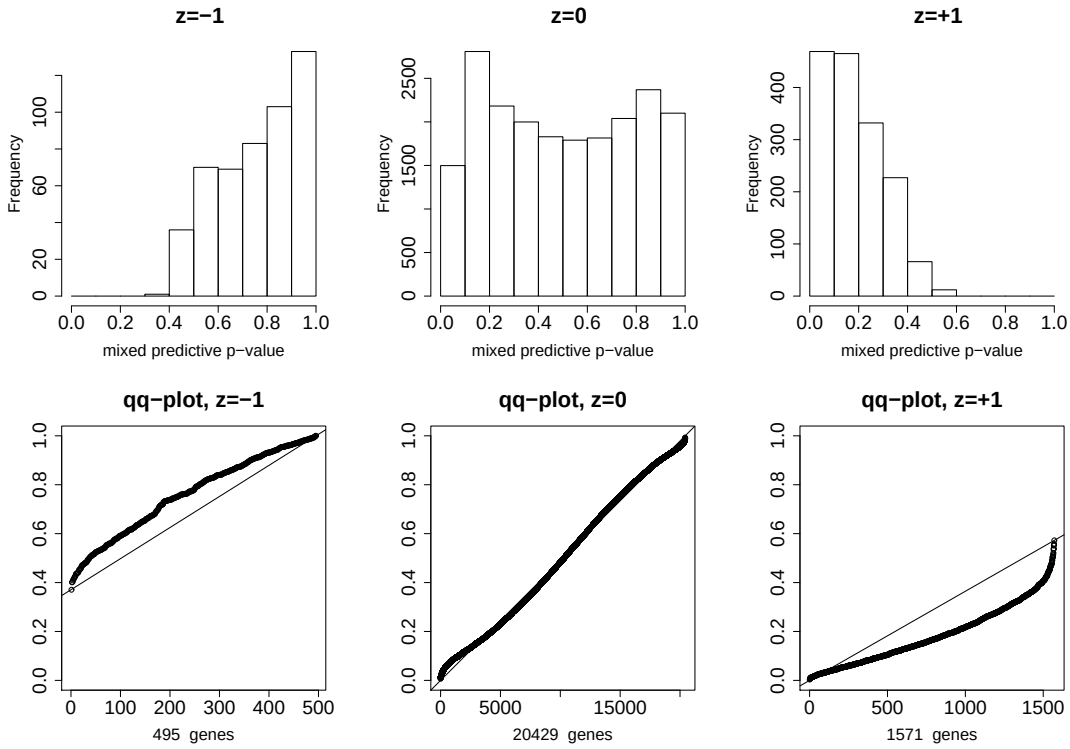

Figure 6. Mixed predictive checks for the mixture of delta-function and Gammas. Mixed predictive p-values for mean differences conditional on z_g (i.e. $\nu_{g,j}$) for $j = -1, 0, 1$. First row shows histograms, second row shows quantile-quantile plots of the distributions versus a Uniform distribution. The histogram and q-q plot for component j is made from genes with $\mathbb{P}(z_g = j|\mathbf{y}^{obs}) > 0.5$. The number of such genes is shown below the q-q plots. Data shown is the wildtype mice, IRS2 data, fitted with the model using the mixture of delta-function and Gammas.

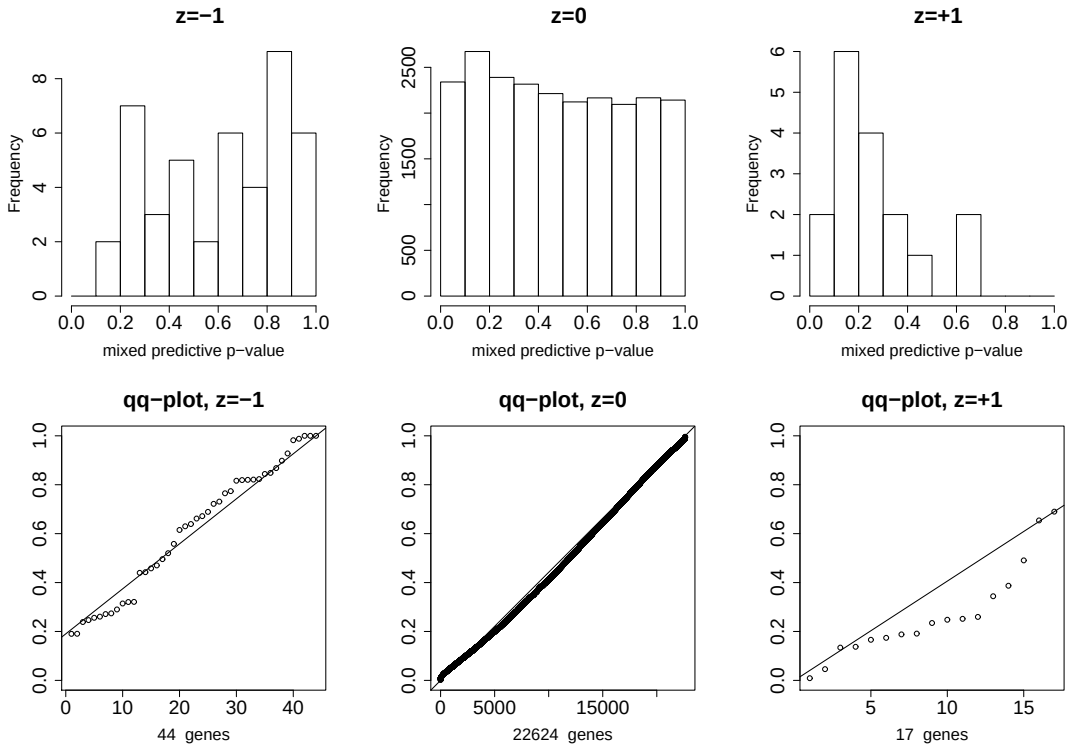

Figure 7. Mixed predictive checks for the mixture of nugget null and Gammas. Mixed predictive p-values for mean differences conditional on z_g (i.e. $\nu_{g,j}$) for $j = -1, 0, 1$. First row shows histograms, second row shows quantile-quantile plots of the distributions versus a Uniform distribution. The histogram and q-q plot for component j is made from genes with $\mathbb{P}(z_g = j|\mathbf{y}^{obs}) > 0.5$. The number of such genes is shown below the q-q plots. Data shown is the wildtype mice, IRS2 data, fitted with the model using the mixture of nugget null and Gammas.

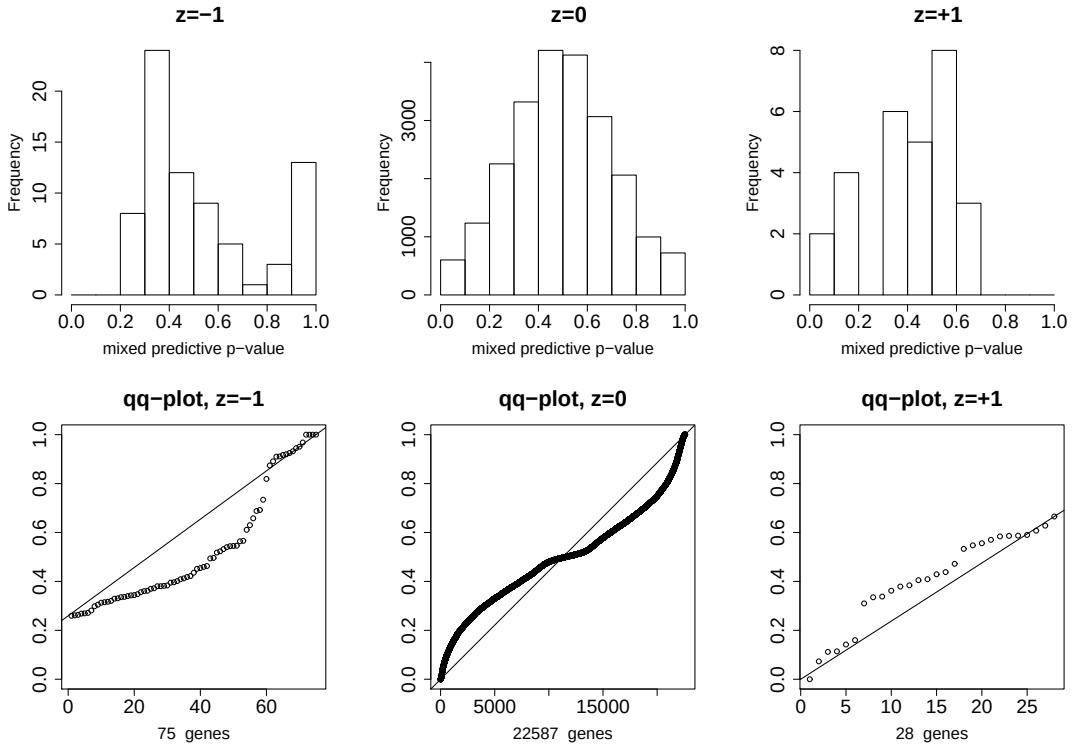

Figure 8. Mixed predictive checks for data processed with the RMA software. Mixed predictive p-values for mean differences conditional on z_g (i.e. $\nu_{g,j}$) for $j = -1, 0, 1$. First row shows histograms, second row shows quantile-quantile plots of the distributions versus a Uniform distribution. The histogram and q-q plot for component j is made from genes with $\mathbb{P}(z_g = j|\mathbf{y}^{obs}) > 0.5$. The number of such genes is shown below the q-q plots. Data shown is the wildtype mice, IRS2 data, processed with the RMA software, fitted with the model using the mixture of delta-function and Gammas.