*Genet. Res., Camb.* (2007), **89**, *pp.* 19–25. © 2007 Cambridge University Press
doi:10.1017/S0016672307008646 Printed in the United Kingdom


# A structural mixed model for variances in differential gene expression studies

FLORENCE JAFFRÉZIC[1]*, GUILLEMETTE MAROT[1], SÉVERINE DEGRELLE[2], ISABELLE HUE[2] AND JEAN-LOUIS FOULLEY[1]
[1] *INRA, UR337 Station de Génétique Quantitative et Appliquée, Jouy-en-Josas 78350, France*
[2] *INRA, UMR 1198; ENVA; CNRS, FRE 2857, Biologie du Développement et Reproduction, Jouy-en-Josas 78350, France*

(*Received 12 March 2007 and in revised form 6 April 2007*)

**Summary**

The importance of variance modelling is now widely known for the analysis of microarray data. In particular the power and accuracy of statistical tests for differential gene expressions are highly dependent on variance modelling. The aim of this paper is to use a structural model on the variances, which includes a condition effect and a random gene effect, and to propose a simple estimation procedure for these parameters by working on the empirical variances. The proposed variance model was compared with various methods on both real and simulated data. It proved to be more powerful than the gene-by-gene analysis and more robust to the number of false positives than the homogeneous variance model. It performed well compared with recently proposed approaches such as SAM and VarMixt even for a small number of replicates, and performed similarly to Limma. The main advantage of the structural model is that, thanks to the use of a linear mixed model on the logarithm of the variances, various factors of variation can easily be incorporated in the model, which is not the case for previously proposed empirical Bayes methods. It is also very fast to compute and is adapted to the comparison of more than two conditions.

## 1. Introduction

Detection of differentially expressed genes relies on statistical tests, typically *t*-tests. A key and critical aspect of these tests is the modelling of the residual variances. The most commonly used approach is to test for differential gene expression one gene at a time. This approach has, in general, low power due to the lack of information on each individual gene (Callow *et al*., 2000). On the other hand, assuming that all the variances are equal and using a common variance estimator can increase the power (Kerr *et al*., 2000) but generates a high rate of false positives when the assumption of homoskedasticity is not true (Cui *et al*., 2005). A number of papers have been devoted to the problem of choosing a suitable variance model for microarray data. In the SAM *t*-test (Tusher *et al*., 2001) a small constant is added to the gene-specific variance estimates in order to stabilize the small variances. Kerr *et al*. (2002) proposed an intensity-dependent variance model where the gene-specific residual variances are modelled as a non-parametric function of the log-intensity. Delmar *et al*. (2005*a*) proposed a mixture model on the gene-variance distributions to identify clusters of genes with equal variances. Cui *et al*. (2005) presented a shrinkage estimator of variance components, using the James–Stein shrinkage concept. Several authors have also proposed hierarchical Bayesian methods, including Lewin *et al*. (2006), Newton *et al*. (2001), Baldi & Long (2001), Lönnstedt & Speed (2002), Wright & Simon (2003), Smyth (2004) and Feng *et al*. (2006).

The aim of this paper is to propose a simple and biologically interpretable model for the variances. The idea is to consider a structural model (Foulley *et al*., 1992) which includes a condition effect and a random gene effect. This model will allow estimation of gene-specific residual variances that will take into account information from all the genes in the data set in a simple and parsimonious way. Two estimation procedures are considered in this paper to estimate the variance parameters: a stochastic approach based

* Corresponding author. e-mail: florence.jaffrezic@jouy.inra.fr

on MCMC techniques and a simple approximate method.

In a simulation study, our method was compared with five other approaches for variance modelling: gene-specific variances, common variance model, SAM (Tusher *et al.*, 2001), VarMixt (Delmar *et al.*, 2005*b*) and Limma (Smyth, 2004). The proposed structural model was also applied to a real functional genomics study on bovine embryos before implantation to find differentially expressed genes according to the reproduction mode, and to a microarray experiment to study the response of the mouse spleen to *in vivo* whole body irradiation.

## 2. Materials and methods

### (i) *Hierarchical model*

Let $y_{ijk}$ be the expression level for gene $i$ ($i=1, \ldots, N$), replicate $j$ ($j=1, \ldots, n_i$) and condition $k$ ($k=1, \ldots, K$). Data are assumed to have been previously normalized. Observations $y_{ijk}$ are modelled with the simple linear model (Delmar *et al.*, 2005*a*):

$$y_{ijk} = m_{ik} + e_{ijk}. \tag{1}$$

The residual terms $e_{ijk}$ are assumed to be independent and normally distributed with mean zero and a variance which can vary both by gene and condition: $e_{ijk} \sim \mathcal{N}(0, \sigma^2_{ik})$.

Estimating one residual variance for each gene within each condition is often not possible due to the lack of replications within each interaction cell. The second step of the proposed hierarchical modelling is therefore to consider a model on the variances that will retain flexibility while keeping the number of parameters reasonably low. As suggested by Foulley *et al.* (1992), a structural model is therefore assumed on the logarithm of the residual variances:

$$\ln(\sigma^2_{ik}) = \mu_k + \delta_{ik}, \tag{2}$$

where $\mu_k$ is a condition effect (assumed fixed) and $\delta_{ik}$ is the gene effect in condition $k$. Here we will assume that the gene effects are independent and normally distributed with mean zero and variance $\tau^2_k$, i.e. $\delta_{ik} \sim \mathcal{N}(0, \tau^2_k)$. Considering the gene effects as random allows us to take into account this source of variation parsimoniously and leads, as shown later, to a shrunk estimator of the variance.

### (ii) *Simple estimation procedure*

Analytical forms of the likelihood function are difficult to obtain in the model presented above, and estimation of the parameters in such a structural model for the variances usually requires the use of stochastic estimation procedures based on MCMC methods. Lewin *et al.* (2006), for example, proposed using Gibbs sampling and estimated the parameters in a Bayesian framework. These stochastic estimation procedures are, however, quite time-consuming due to the large number of simulations required to obtain accurate estimates of the parameters.

Here we propose a simple and efficient approximate method to obtain estimates of the parameters in the structural model for the variances. These estimates were compared with those obtained with Gibbs sampling using the software WINBUGS (Spiegelhalter *et al.*, 2004). The idea of the proposed estimation procedure is to base inference of the variance parameters on the empirical variances.

For each gene $i$, let $s^2_{ik}$ be the empirical variance defined as

$$s^2_{ik} = \frac{1}{n_{ik}-1} \sum_{j=1}^{n_{ik}} (y_{ijk} - y_{ik.})^2, \tag{3}$$

where $y_{ijk}$ represents the expression level for replicate $j$ of gene $i$ in condition $k$. Let $y_{ik.}$ be the average expression level for gene $i$ over all replicates in condition $k$: $y_{ik.} = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} y_{ijk}$. For the proposed estimation procedure, the structural model is assumed on the logarithm of the empirical variances:

$$\ln(s^2_{ik}) = \mu_k + \delta_{ik} + \varepsilon_{ik}, \tag{4}$$

where $\varepsilon_{ik}$ is a sampling error due to the estimation of the true variances $\sigma^2_{ik}$ by the empirical variances $s^2_{ik}$. Residuals $\varepsilon_{ik}$ are assumed independent and normally distributed with mean zero and variance $\omega^2_{ik}$: $\varepsilon_{ik} \sim \mathcal{N}(0, \omega^2_{ik})$. According to the asymptotic theory (Layard, 1973), the sampling variances $\omega^2_{ik}$ can be estimated by $\omega^2_{ik} = 2/d_{ik}$, where $d_{ik}$ corresponds to the degrees of freedom for gene $i$ in condition $k$. Usually $d_{ik} = n_{ik} - 1$, where $n_{ik}$ represents the number of replicates for gene $i$ in condition $k$. As previously, $\delta_{ik}$ is assumed to be a random gene effect in condition $k$: $\delta_{ik} \sim \mathcal{N}(0, \tau^2_k)$, and $\mu_k$ is a fixed effect which represents the condition effect. Both parameters $\tau^2_k$ and $\mu_k$ can be estimated by classical linear mixed model estimation procedures.

Due to the use of normal conjugate distributions – $\ln s^2_{ik} | \ln \sigma^2_{ik} \sim \mathcal{N}(\ln \sigma^2_{ik}, \omega^2_{ik})$ and $\ln \sigma^2_{ik} \sim \mathcal{N}(\mu_k, \tau^2_k)$ – it follows that the best predictor of $\ln \sigma^2_{ik}$ is

$$\widehat{\ln \sigma^2_{ik}} = \mu_k + \lambda_{ik}(\ln s^2_{ik} - \mu_k), \tag{5}$$

where $\lambda_{ik} = \tau^2_k / (\tau^2_k + \omega^2_{ik})$ is a shrinkage factor of $\ln \sigma^2_{ik}$ towards $\mu_k$. When parameters $\tau^2_k$ tend to zero, we obtain a pooled estimator and a common variance for all genes within each condition: $\hat{\mu}_k = \sum_i (d_{ik} \ln s^2_{ik}) / \sum_i d_{ik}$. On the other hand, if parameters $\tau^2_k$ tend to infinity, the shrinkage factors $\lambda_{ik}$ become 1. There is no shrinkage, and one variance is estimated for each gene in each condition as: $\widehat{\ln \sigma^2_{ik}} = \ln s^2_{ik}$.

https://doi.org/10.1017/S0016672307008646 Published online by Cambridge University Press

(iii) *Degrees of freedom of the T statistic*

To test whether gene $i$ is differentially expressed between condition $k$ and condition $l$ the test statistic is

$$t_{i,kl} = \frac{m_{ik} - m_{il}}{\sqrt{\hat{\sigma}^2_{ik}/n_{ik} + \hat{\sigma}^2_{il}/n_{il}}}, \quad (6)$$

where $\hat{\sigma}^2_{ik}$ and $\hat{\sigma}^2_{il}$ are estimations under the proposed structural model presented above. The exact distribution of this test statistic under the null hypothesis is unknown and determination of the $p$ values can be obtained by permutations. As pointed out by Cui *et al.* (2005) permutations are, however, very time-consuming, especially when a large number of genes are analysed. To obtain a fast and efficient procedure, we therefore propose considering an approximate Student distribution. In fact, under the structural model, the test statistic corresponds to the so-called Welch's statistic which follows approximately a Student distribution (Moser & Stevens, 1992) with $\nu_i$ degrees of freedom. For each gene, we propose calculating the degrees of freedom of the $T$ statistic by the classically used Satterthwaite's method, as follows:

$$\nu_i = \frac{2(\hat{\sigma}^2_{ik} + \hat{\sigma}^2_{il})^2}{\mathrm{Var}(\hat{\sigma}^2_{ik}) + \mathrm{Var}(\hat{\sigma}^2_{il})}, \quad (7)$$

where $\hat{\sigma}^2_{ik}$ and $\hat{\sigma}^2_{il}$ are the variance parameter estimations obtained with the structural model and the variances of these estimations can be calculated as: $\mathrm{Var}(\hat{\sigma}^2_{ik}) = (\hat{\sigma}^2_{ik})^2 \mathrm{Var}(\widehat{\ln \sigma^2_{ik}})$, where $\mathrm{Var}(\widehat{\ln \sigma^2_{ik}}) \approx (1/\tau^2_k + d_{ik}/2)^{-1}$ with $d_{ik} = (n_{ik} - 1)$ for condition $k$.

An R function 'SMVar' implementing the structural model for the detection of differentially expressed genes is available upon request from the first or second author.

## 3. Application

The proposed structural model was applied here to two sets of real data to find differentially expressed genes in bovine embryos according to the reproductive mode and in mice to study the spleen response to irradiation.

(i) *Reproductive mode in bovine embryos*

(a) *Presentation of the data.* This variance modelling was applied to a functional genomics study on bovine embryos before implantation. The experimental protocol is described in detail by Degrelle (2006). The aim of this study was to find differentially expressed genes in the embryos according to the reproductive mode. Three reproductive modes were investigated: artificial insemination (AI), *in vitro* fertilization (IVF) and cloning (somatic cell nuclear transfer, SCNT). Three different lines of clones were studied. They were established from ear skin biopsies of three Holstein heifers. In total, 10 Holstein embryos were available for AI, IVF and each of the three lines of clones. In total, 10 214 unique cDNA were spotted onto Nylon N+ membranes (Amersham Biosciences) at the CRB GADIE platform (INRA, Jouy-en-Josas). The bovine 10K array will be fully described in a forthcoming paper (Degrelle *et al.*, unpublished). For each embryo ($n = 50$), RNA was isolated, amplified (MessageAmp aRNA Kit, Ambion) and hybridized onto the array. The membranes were exposed to phosphor screens for 7 days. The hybridization signals were quantified using Imagene 5.5 software (BioDiscovery) on the PICT platform (INRA, Jouy-en-Josas). Gene expression data were $\log_2$ transformed. Data were centred by membrane and by gene. No further normalization was needed on this bovine data set.

(b) *Variance parameter estimations.* For the structural model, the list of differentially expressed genes found with the approximated estimation method was compared with the list obtained with the exact MCMC estimations using Gibbs sampling with WINBUGS software (Spiegelhalter *et al.*, 2004). As the posterior distributions of the variance parameters were highly asymmetrical, we chose the posterior mode with a uniform prior on the standard deviations (Gelman, 2005) as a point estimate of the variance parameters, which is close to the REML estimation of the variance parameters. The structural model was compared with the mixture model approaches proposed by Delmar *et al.* (2005*b*): VM and VM2. In VM2, each gene is assigned to one of the groups of homogeneous variance determined by the mixture model. VM is more flexible as it does a partial assignment of genes to variance groups, taking into account the probabilities of belonging to each group. Classical methods such as Limma (Smyth, 2004), SAM (Tusher *et al.*, 2001), gene-by-gene analysis and the homogeneous variance model were also applied to this data set. To make each method comparable, a Benjamini & Hochberg (1995) correction (BH correction) was performed on the raw $p$ values to correct for multiple tests.

The proposed structural model is similar in spirit to that of Baldi & Long (2001), except that the use of log-normal distributions instead of Gamma gives the possibility of directly estimating the shrinkage parameter, which is a crucial parameter for the variance estimations, whereas it has to be specified *a priori* by the user in Cyber-T. Moreover, the structural model allows the easy incorporation of factors of variation other than the gene and condition effects. Analyses performed here will therefore not be

https://doi.org/10.1017/S0016672307008646 Published online by Cambridge University Press

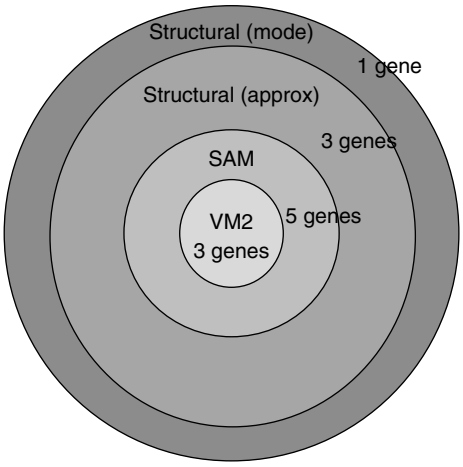


Fig. 1. Venn diagram for the differentially expressed genes detected at a 10% BH threshold in the real bovine data set with four methods: structural model using the posterior mode in the Gibbs sampling estimations, structural model with the approximate method, SAM and VM2.

compared with the method presented by Baldi & Long (2001).

(c) *Results*. The Venn diagram presented in Fig. 1 illustrates the list of differentially expressed genes found with the different methods at a 10% BH threshold. With the structural model and the proposed approximate estimation method, 11 genes were found which were all included in the 12 genes found with Gibbs sampling estimations using posterior mode estimates.

The SAM approach detected 8 genes at 10% which were all included in the 11 genes detected with the structural model. In this analysis, VM and VM2 were found to lack power as no gene was detected with VM and only 3 genes were detected with VM2. This may be due to the fact that the VarMixt methods were designed for comparing only two conditions as the variance of the gene expression difference is modelled, whereas in this example five conditions were compared. In contrast, the structural approach models the variances in each condition and can therefore readily be applied to the comparison of more than two conditions. At a 30% BH threshold, 9 genes were detected with VM and 6 with VM2. All of them were included in the 11 genes detected with the structural model. Similarly, Limma detected only 1 differentially expressed gene at a BH threshold of 10% as well as 30%.

The homogeneous variance model found far more genes than other methods, but a histogram of the $p$ values showed that the assumption of a common variance is not appropriate for these data, as shown in Fig. 2. In fact, the distribution of the $p$ values was not uniform under the null hypothesis. It is therefore expected that a large proportion of the detected genes are false positives.

(ii) *Mouse spleen data*

(a) *Presentation of the data*. These data were presented and analysed by Delmar *et al*. (2005*a*), and are publicly available in the R VarMixt package (Delmar *et al*., 2005*b*). The goal of this experiment was to study the response of the mouse spleen to *in vivo* whole-body irradiation. Experimental data were generated with two-colour complementary DNA microarray assays comparing the spleen of irradiated (treated) and normal (control) mice. The data consist of three dye-swaps. The 'treated' samples were obtained from three independent mice (one mouse per swap) 3 hours after irradiation at 1 Gy. The 'control' sample was obtained from pooling several normal mice. The same control sample was used in all the hybridization experiments. There are 4360 genes in each array. Composition of the arrays is described in Preisser *et al*. (2004). Data were previously normalized as described by Delmar *et al*. (2005*a*).

(b) *Results*. Three methods have been applied to find differentially expressed genes in these data, namely Limma (Smyth, 2004), VM (Delmar *et al*., 2005*b*) and the structural model proposed here. The Benjamini & Hochberg (1995) procedure at a 5% threshold was used to correct for multiple tests. In total 112 genes were detected with Limma, 113 with VM and 125 with the structural model. Among them, 104 genes were found by all three methods, as shown in the Venn diagram in Fig. 3.

## 4. Simulation study

A simulation study was performed to compare the proposed structural model with the variance modelling implemented in Limma (Smyth, 2004), SAM (Tusher *et al*., 2001) and VarMixt (Delmar *et al*., 2005*b*), as well as with the simple gene-by-gene analyis and homogenenous variance model. In the first simulation, paired data were studied from the 'two-colour' experiment in mice presented by Delmar *et al*. (2005*a*) and analysed in the previous section. The second simulation study is based on the real bovine data presented above; these are therefore unpaired data. For each of the methods, a BH correction was performed on the raw $p$ values to account for multiple tests.

(i) *Simulation 1*

(a) *Data*. The first simulation was performed with the same parameters as used by Delmar *et al*.

https://doi.org/10.1017/S0016672307008646 Published online by Cambridge University Press

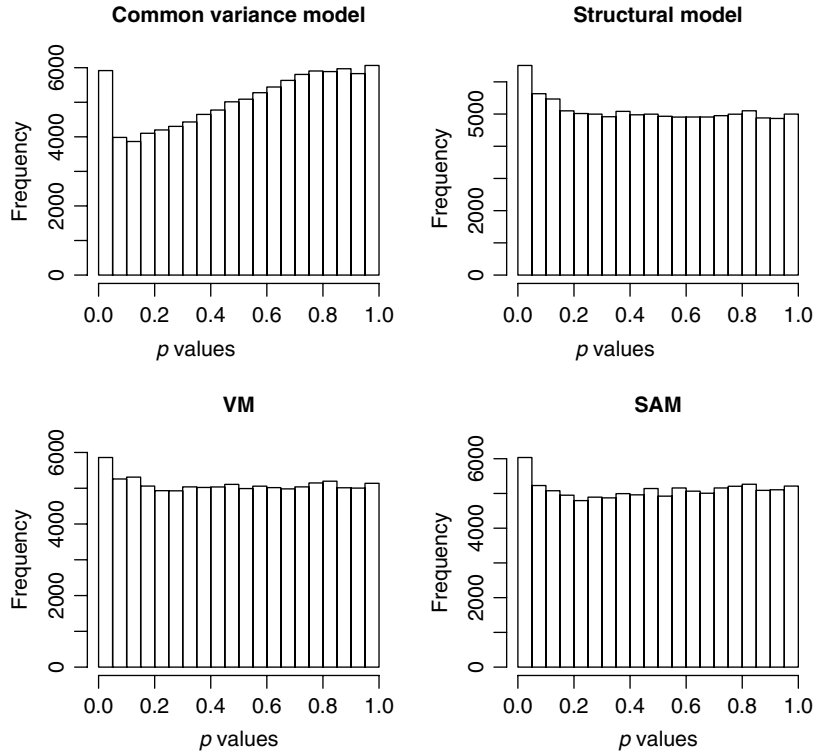


Fig. 2. Histogram of the raw $p$ values for the real bovine data analysis with four different models: the common variance model, the proposed structural model, VM and SAM.

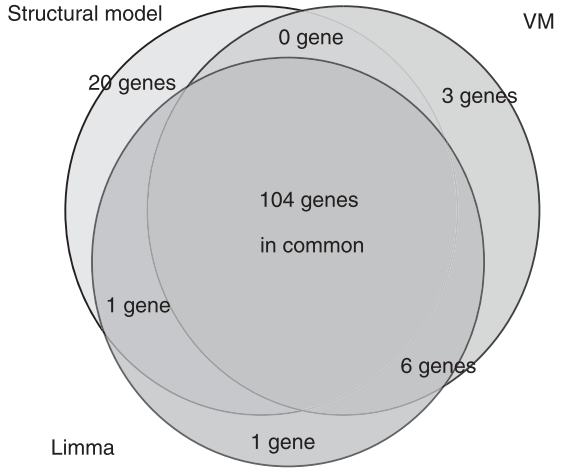


Fig. 3. Venn diagram for the differentially expressed genes detected at a 5% BH threshold in the real mouse data set with three methods: structural model, VM and Limma.

(2005*a*), which were obtained from the real mouse data set analysed above. The simulated data had 4360 genes which were assumed normally distributed. One per cent of the genes were simulated with a non-zero mean log-ratio. These 43 genes were simulated with a mean log-ratio ranging uniformly from 0.25 to 0.9. Gene variances were estimated from the real data by a gene-by-gene analysis and were randomly assigned to the differentially expressed genes in each simulated data set as in Delmar *et al*. (2005*a*).

(b) *Model fitting*. In this data set, each array is hybridized with both a control and a treated sample. Therefore, for each gene the two observations from the same array were treated as paired data. Each model was fitted on the logarithm of the ratio of observed intensity (log-ratio). Let $y_{ij}$ be the log-ratio for gene $i$ in replicate $j$. It is modelled by

$$y_{ij} = m_i + e_{ij}, \tag{8}$$

where $e_{ij} \sim \mathcal{N}(0, \sigma_i^2)$. For the structural model, the residual variances are now modelled as: $\ln \sigma_i^2 = \mu + \delta_i$, where $\delta_i \sim \mathcal{N}(0, \tau^2)$. For these paired data, the measure of differential expression for gene $i$ between the two conditions is now defined as the mean log-ratio for gene $i$:

$$\Delta_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}. \tag{9}$$

In this first simulation study, the paired Limma, VM, VM2, SAM, gene-by-gene and homoskedastic models were also used. The results were averaged over 100 simulated data sets and are presented in Table 1 for a 5% BH threshold.

(c) *Results*. It was found that for relatively large numbers of replicates (eight or more), all four methods (Limma, SAM, VarMixt and structural model) perform quite well. The homogeneous model,

https://doi.org/10.1017/S0016672307008646 Published online by Cambridge University Press

Table 1. *Results of the simulations based on the mouse paired data set at a 5% BH threshold*

| | No. of replicates[a] | | |
|---|---|---|---|
| | 5 | 8 | 10 |
| *No. of true positives* | | | |
| Structural model | 29.4 (3.1) | 39.8 (1.6) | 41.4 (1.2) |
| VM | 33.7 (2.4) | 40.2 (1.4) | 41.6 (1.1) |
| VM2 | 32.5 (2.7) | 39.7 (1.7) | 41.3 (1.3) |
| SAM | 0.0 (0.0) | 39.9 (1.7) | 40.9 (1.5) |
| Limma | 33.0 (2.5) | 40.0 (1.6) | 41.4 (1.2) |
| Gene-specific | 13.8 (4.2) | 37.1 (2.2) | 39.9 (1.7) |
| Homoskedastic | 39.9 (1.4) | 42.4 (0.8) | 42.8 (0.5) |
| *No. of false positives* | | | |
| Structural model | 1.7 (1.5) | 2.0 (1.3) | 2.2 (1.8) |
| VM | 2.0 (1.6) | 2.3 (1.6) | 2.4 (1.8) |
| VM2 | 2.0 (1.7) | 2.2 (1.4) | 2.1 (1.7) |
| SAM | 0.0 (0.0) | 1.7 (1.5) | 2.1 (1.7) |
| Limma | 0.0 (0.0) | 0.0 (0.0) | 0.0 (0.0) |
| Gene-specific | 0.8 (1.0) | 1.9 (1.2) | 2.4 (1.8) |
| Homoskedastic | 49.8 (7.9) | 50.0 (7.9) | 51.7 (8.7) |

Values are the mean (SD) over 100 simulations.
[a] Replicates correspond to the number of measurements for each gene within each condition.

Table 2. *Results of the simulations based on a subset of the bovine reproductive data set at a 10% BH threshold*

| | No. of replicates[a] | | |
|---|---|---|---|
| | 5 | 8 | 10 |
| *No. of true positives* | | | |
| Structural model | 18.7 (6.6) | 53.9 (5.6) | 56.2 (4.6) |
| VM | 10.4 (5.5) | 47.7 (5.2) | 60.2 (4.7) |
| VM2 | 6.4 (5.1) | 43.1 (5.5) | 58.2 (4.8) |
| SAM | 11.5 (5.5) | 50.6 (5.5) | 63.3 (4.8) |
| Limma | 17.8 (6.0) | 49.3 (5.0) | 61.2 (4.4) |
| Gene-specific | 6.2 (4.1) | 39.2 (5.0) | 54.1 (4.4) |
| Homoskedastic | 37.5 (4.43) | 63.2 (4.4) | 73.5 (3.9) |
| *No. of false positives* | | | |
| Structural model | 6.2 (4.2) | 14.8 (5.3) | 16.8 (5.1) |
| VM | 2.0 (1.9) | 8.7 (3.7) | 11.6 (4.0) |
| VM2 | 1.3 (1.7) | 7.3 (3.6) | 10.3 (3.7) |
| SAM | 2.3 (2.2) | 11.2 (4.5) | 13.7 (4.6) |
| Limma | 5.0 (3.4) | 14.0 (4.9) | 16.8 (5.0) |
| Gene-specific | 1.2 (1.5) | 7.9 (3.6) | 11.3 (4.3) |
| Homoskedastic | 89.1 (11.7) | 102.1 (10.6) | 104.5 (11.2) |

Values are the mean (SD) over 100 simulations.
[a] Replicates correspond to the number of measurements for each gene within each condition.

however, had a very large rate of false positives, which shows that this assumption is here highly unrealistic. The structural model still performs quite well for fewer replicates (five replicates), even with the approximate estimation method based on empirical variances. As already observed by Delmar *et al.* (2005*a*), however, the paired SAM method performs very poorly for five replicates as it detects no differentially expressed genes. As expected, the gene-by-gene analysis showed a lack of power with this small number of replicates and the homogeneous variance model still had a large number of false positives.

(ii) *Simulation 2*

(a) *Data*. This second simulation was based on the parameters estimated on the real data set presented above on bovine embryos. Due to the required computing time, only three conditions among the five were considered and 5000 genes among the 10 214. Among them, 100 genes were simulated to be differentially expressed for one of the conditions compared with the two others. For these genes, the mean log-ratio was simulated according to a Gamma (10.8,0.07). These parameters were determined from the real data set. Gene variances used for the simulations were estimated from the real data by a gene-by-gene analysis and were randomly distributed within the set of differentially expressed genes for each simulation. In this study, only unpaired methods were used as the real data came from a membrane experiment and not a 'two-colour' experiment. Results were averaged over 100 simulated data sets and are presented in Table 2 for a 10% BH threshold.

(b) *Results*. In the case of the comparison of more than two conditions, as already observed in the real data analysis, the structural model had more power than VM and SAM, especially in the case of a small number of replicates (five replicates here). In fact, 19 true positives were detected on average with the structural model at a 10% BH threshold, whereas fewer than seven genes were detected with VM2, fewer than 11 with VM and 12 with SAM. This is due to the fact that the structural approach models directly the variance of each gene within each condition, whereas the VarMixt methods model the variance of the difference in gene expression in two conditions. On the other hand, the Limma approach also works quite well in this case with 18 true positives detected. In the case of 10 replicates the same pattern is oberved, although the differences between methods are slightly smaller than for five replicates.

## 5. Discussion

The first simulation study showed that the proposed structural model for paired data performed similarly to the VarMixt approach. The paired SAM method, however, showed a considerable lack of power in this

https://doi.org/10.1017/S0016672307008646 Published online by Cambridge University Press

analysis when the number of replicates was small. As expected, the structural model clearly outperformed the homogeneous variance model and the gene-by-gene analysis. The proposed approximate estimation procedure, based on empirical variances, still performed well for a small number of replicates (five replicates).

In the first real data analysis and the second simulation study, the structural model was found to be more powerful than VM, VM2 and SAM. This was due to the fact that more than two conditions were compared whereas VM and VM2 were initially developed for the comparison of only two conditions. In fact, the mixture model is based directly on the variance of the gene expression difference instead of modelling the variance in each condition.

The structural model was found here to perform similarly to the Limma approach (Smyth, 2004). The main advantage of the structural model is, however, that the use of a linear mixed model on the log of the variances provides a larger modelling flexibility. In fact, here a condition and gene effects were included in the model, but it could easily be extended to other mixed models including, for example, a sex effect or even functions of time. This is much more difficult to achieve when considering an inverse chi-square distribution on the variances, as proposed by Smyth (2004).

## References


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

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **85**, 289–300.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. & Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in hdl-deficient mice. *Genome Research* **10**, 2022–2029.

Cui, X., Gene Hwang, J. T., Qiu, J., Blades, N. J. & Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59–75.

Degrelle, S. (2006). Croissance et différenciation du trophoblaste de mammifères en début de gestation: étude par génomique fonctionnelle de l'embryon bovin normal et cloné. PhD thesis, Université de Versailles-Saint-Quentin-en-Yvelines.

Delmar, P., Robin, S., Tronik-Le Roux, D. & Daudin, J.-J. (2005*a*). Mixture model on the variance for the differential analysis of gene expression data. *Applied Statistics* **54**, 31–50.

Delmar, P., Robin, S. & Daudin, J.-J. (2005*b*). VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics* **21**, 502–508.

Feng, S., Wolfinger, R. D., Chu, T. M., Gibson, G. C. & McGraw, L. A. (2006). Empirical Bayes analysis of variance component models for microarray data. *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 197–209.

Foulley, J.-L., San Cristobal, M., Gianola, D. & Im, S. (1992). Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. *Computational Statistics and Data Analysis* **13**, 291–305.

Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 1–19.

Kerr, M. K., Martin, M. & Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

Kerr, M., Afshari, C., Bennett, L., Bushel, P., Martinez, J., Walker, N. & Churchill, G. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* **12**, 203–217.

Layard, M. W. J. (1973). Robust large-sample tests for homogeneity of variances. *Journal of the American Statistical Association* **68**, 195–198.

Lewin, A., Richardson, S., Marshall, C., Glazier, A. & Aitman, T. (2006). Bayesian modeling of differential gene expression. *Biometrics* **62**, 1–9.

Lönnstedt, I. & Speed, T. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

Moser, B. K. & Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *The American Statistician* **46**, 19–21.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. & Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

Preisser, L., Houot, L., Teillet, L., Kortulewski, T., Morel, A., Tronik-Le Roux, D. & Corman, B. (2004). Gene expression in aging kidney and pituitary. *Biogerontology* **5**, 39–47.

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., Thomas, A. & Best, N. G. (2004). *WinBUGS Version 1.4 User Manual*. Cambridge: Medical Research Council Biostatistics Unit. Available from http://www.mrc-bsu.cam.ac.uk/bugs

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

Wright, G. W. & Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448–2455.