Statistical Modelling 2008; **8**(1): 23–39

Simultaneous inference for multiple testing and clustering via a Dirichlet process mixture model

David B Dahl[1], Qianxing Mo[2] and Marina Vannucci[3]
[1] Texas A&M University, US
[2] Memorial Sloan-Kettering Cancer Center, US
[3] Rice University, US

Abstract: We propose a Bayesian nonparametric regression model that exploits clustering for increased sensitivity in multiple hypothesis testing. We build on the recently proposed BEMMA (Bayesian Effects Models for Microarrays) method which is able to model dependence among objects through clustering and then estimates hypothesis-testing parameters averaged over clustering uncertainty. We propose several improvements. First, we separate the clustering of the regression coefficients from the part of the model that accommodates heteroscedasticity. Second, our model accommodates a wider class of experimental designs, such as permitting covariates and not requiring independent sampling. Third, we provide a more satisfactory treatment of nuisance parameters and some hyperparameters. Finally, we do not require the arbitrary designation of a reference treatment. The proposed method is compared in a simulation study to ANOVA and the BEMMA methods.

Key words: Bayesian nonparametrics; correlated hypothesis tests; model-based clustering; multiple comparisons

Received October 2006; **revised** January 2007; **accepted** January 2007

1 Introduction

Clustering and multiple hypothesis testing are two active areas of research in high dimensional statistical inference. The multiple comparisons problem occurs when subjecting objects to the same rejection criterion which, for example, attempts to detect a shift in the distribution of an object's data due to different treatment conditions. Statistical dependence among objects should be accommodated in multiple testing, although this is not of primary interest. Conversely, the goal of clustering is to make inference on the statistical dependence among objects.

The two inferential tasks of clustering and multiple testing are typically treated separately. Recently, however, Dahl and Newton (2007), Yuan and Kendziorski (2006), and Tibshirani and Wasserman (2006) proposed hybrid methods, all set in the context of microarrays, that attempt to exploit dependence among objects (for example, genes) to help in multiple testing (for example, detecting differentially

Address for correspondence: David B Dahl, Department of Statistics, Texas A&M University, College Station, TX 77843. E-mail: dahl@state.tamu.edu

© 2008 SAGE Publications 10.1177/1471082X0700800103

expressed genes). The BEMMA method of Dahl and Newton (2007) uses a Dirichlet process mixture model to estimate parameters capturing evidence for differential expression in a model-based clustering procedure that averages over clustering uncertainty. Yuan and Kendziorzki (2006) use a finite mixture model assigning genes to clusters and discrete expression patterns. The method of Tibshirani and Wasserman (2006) averages univariate scores (for example, t-test statistics) of highly correlated genes. This latter method is very straight-forward but information about the dependence among genes may be lost when computing the univariate scores.

We propose a hybrid method for simultaneous inference on multiple testing and clustering. We name our proposal SIMTAC, an acronym for 'Simultaneous Inference for Multiple Testing And Clustering'. Our method provides several extensions to the BEMMA method of Dahl and Newton (2007). First, we form clusters on the regression coefficients and accommodate heteroscedasticity using two independent Dirichlet process (DP) priors (rather than one DP prior as in Dahl and Newton, 2007). Second, our model permits an arbitrary experimental design matrix (for example, containing covariates) and does not require independent sampling. Third, we handle the object-specific shifts in a natural Bayesian fashion and are more flexible in our treatment of the mass parameters of the DP priors. Finally, unlike Dahl and Newton (2007), we do not need to specify a reference treatment.

In the remainder of the paper, we first describe the SIMTAC method and then describe how inference on clustering and hypothesis testing can be conducted. We end the paper with a simulation study that compares our method with Analysis of Variance (ANOVA) and the BEMMA method of Dahl and Newton (2007).

2 Model

2.1 Sampling distribution

Suppose there are K observations on each of G objects. For each object g, our SIMTAC method assumes that the data vector $\boldsymbol{d}_g$ has the following K-dimensional multivariate normal distribution:

$$\boldsymbol{d}_g \mid \mu_g, \boldsymbol{\beta}_g, \lambda_g \sim \mathrm{N}_K\left(\boldsymbol{d}_g \mid \mu_g \boldsymbol{j} + \mathbf{X}\boldsymbol{\beta}_g, \lambda_g \mathbf{M}\right), \tag{2.1}$$

where μ_g is an object-specific mean, $\boldsymbol{j}$ is a vector of ones, $\mathbf{X}$ is a $K \times L$ design matrix, $\boldsymbol{\beta}_g$ is a vector of L object-specific regression coefficients, $\mathbf{M}$ is the inverse of a correlation matrix of the K observations from an object, and λ_g is an object-specific precision (that is, inverse of the variance).

The goals of our data modelling are two-fold: (1) infer the clustering of objects, where objects within a cluster share a common value for their regression coefficients, and (2) test a hypothesis for each object regarding its regression coefficient. The precise form of the hypothesis will depend on the experiment design and objectives. In a two-treatment setting, for example, we would test whether $\boldsymbol{\beta}_g = 0$. We give other examples in Section 3.2.

Dahl and Newton (2007) proposed a similar model for the analysis of microarray data. In this context, G would be the number of genes, K would be the number of microarrays, and $\boldsymbol{d}_g$ would be the suitably-transformed expression data of gene g. The sampling distribution in (2.1), however, is more flexible than that of Dahl and Newton (2007). While the SIMTAC method permits arbitrary $\mathbf{X}$ and $\mathbf{M}$ matrices, Dahl and Newton (2007) assumed an ANOVA setting (for example, no covariates) and independence among observations within an object (that is, in our notation, they set $\mathbf{M}$ to an identity matrix).

Note that $\mathbf{X}$ and $\mathbf{M}$ are known and common to all objects, whereas μ_g, $\boldsymbol{\beta}_g$ and λ_g are unknown object-specific parameters. Since an intercept is explicit through $\mu_g \boldsymbol{j}$, the design matrix $\mathbf{X}$ does not contain a column vector of ones. For example, in a two-treatment experiment, $\mathbf{X}$ would have just one column containing the dummy variable differentiating the two treatments. Of course, in addition to dummy variables for treatments, $\mathbf{X}$ can also contain covariates. For experimental designs involving independent sampling (for example, the typical time-course microarray experiment in which subjects are sacrificed rather than providing repeated measures), $\mathbf{M}$ is simply the identity matrix. In the case where $\mathbf{M}$ is not known, a Wishart prior distribution can be used and $\mathbf{M}$ can be estimated from its posterior distribution. This approach is conceptually straightforward to implement but is, of course, computationally more demanding than fixing the value of $\mathbf{M}$.

2.2 Clusterings via Dirichlet process priors

Our SIMTAC method makes use of DP (Ferguson, 1973) methodology and is thus a Dirichlet process mixture (DPM) model. See Müller and Quintana (2004) for a review. We achieve simultaneous inference on clustering and hypothesis testing by exploiting the fact that realizations of a DP are almost-surely discrete, random distributions.

We place a DP prior (Antoniak, 1974) on the regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$, thereby inducing ties among their values. That is, for every pair of object $i \neq j$, there is positive probability that $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$. Two objects $i \neq j$ are said to be clustered in terms of their regression coefficients if and only if $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$. The clustering of the objects encoded by the ties of the regression coefficients will simply be referred to as the 'clustering of the regression coefficients', although it should be understood that it is the data themselves that are clustered.

A set partition parameterization is helpful throughout the paper. A set partition $\boldsymbol{\pi} = \{S_1, \ldots, S_q\}$ of $S_0 = \{1, \ldots, G\}$ satisfies $\cup_{S \in \boldsymbol{\pi}} S = S_0$, $S \cap S^* = \emptyset$ for all $S \neq S^*$, and $S \neq \emptyset$ for all $S \in \boldsymbol{\pi}$. A cluster $S \in \boldsymbol{\pi}$ for regression coefficients is a set of indices such that, for all $i \neq j \in S$, $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$. Let $\boldsymbol{\beta}_S$ denote the common value of the regression coefficients corresponding to cluster S. Using this set partition notation, the regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ can be reparameterized as a partition $\boldsymbol{\pi}_{\boldsymbol{\beta}}$ and a collection of unique model parameters $\boldsymbol{\phi}_{\boldsymbol{\beta}} = (\boldsymbol{\beta}_{S_1}, \ldots, \boldsymbol{\beta}_{S_q})$. In this paper, the terms clustering and set partition are used interchangeably.

To accommodate heteroscedasticity among the objects, the precisions $\lambda_1, \ldots, \lambda_G$ form another clustering through the use of a separate DP prior on them. Let $\boldsymbol{\pi}_\lambda$ and $\boldsymbol{\phi}_\lambda = (\lambda_{S_1}, \ldots, \lambda_{S_q})$ be the set partition parameterization of the precisions $\lambda_1, \ldots, \lambda_G$. Thus, the SIMTAC method entails two clusterings: one based on ties among the regression coefficients and another based on ties among the precisions. The two clusterings will likely have distinct configurations and numbers of clusters.

2.3 Prior distribution

The prior specification is completed by choosing a standard conjugate prior for the means $\mu_1, \ldots, \mu_G$. In all, the joint prior distribution is:

$$\begin{aligned} \mu_g &\sim \mathrm{N}\left(\mu_g \mid m_\mu, p_\mu\right) \\ \boldsymbol{\beta}_g \mid G_{\boldsymbol{\beta}} &\sim G_{\boldsymbol{\beta}} \\ G_{\boldsymbol{\beta}} &\sim \mathrm{DP}\left(\alpha_{\boldsymbol{\beta}} G^{\star}_{\boldsymbol{\beta}}\right) \\ \lambda_g \mid G_\lambda &\sim G_\lambda \\ G_\lambda &\sim \mathrm{DP}\left(\alpha_\lambda G^{\star}_\lambda\right). \end{aligned}$$

The centring distributions in the DP priors are $G^{\star}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \mathrm{N}_L(\boldsymbol{\beta} \mid \boldsymbol{m}_{\boldsymbol{\beta}}, \mathbf{P}_{\boldsymbol{\beta}})$ and $G^{\star}_\lambda(\lambda) = \mathrm{Ga}(\lambda \mid a_\lambda, b_\lambda)$ having mean a_λ / b_λ. Typically $\boldsymbol{m}_{\boldsymbol{\beta}}$ is set to be the zero vector. One could specify the other hyperparameters $m_\mu, p_\mu, \mathbf{P}_{\boldsymbol{\beta}}, a_\lambda, b_\lambda$ either based on prior belief or using the empirical Bayes approach in Appendix A. Our experience is that the empirical Bayes approach works well in practice and that testing and clustering results are robust to departures from these recommendations. For example, in our simulation study, we initially made the mistake of treating $\mathbf{P}_{\boldsymbol{\beta}}$ as a variance instead of as a precision, yet the testing and clustering results were nearly unchanged from the correct results that we report in Section 4.

The mass parameters $\alpha_{\boldsymbol{\beta}}$ and α_λ influence the number of clusters; values close to zero induce many ties and larger values induce fewer ties. One approach is to set them to constant values based on prior experience. Dahl and Newton (2007) used an empirical Bayes approach to fix their values. We, instead, allow for greater flexibility by placing two independent priors of the mass parameters, each of the form described by Escobar and West (1995). In any case, quantities of interest are often robust to the treatment of the mass parameters (for example, Medvedovic and Sivaganesan (2002) and Dahl and Newton (2007)).

A major difference between our SIMTAC method and the BEMMA method of Dahl and Newton (2007) is in the clustering of the regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ and the precisions $\lambda_1, \ldots, \lambda_G$. BEMMA has one simultaneous clustering for the joint object-specific parameters $(\boldsymbol{\beta}_1, \lambda_1), \ldots, (\boldsymbol{\beta}_G, \lambda_G)$, each encapsulating both the regression coefficients and the precision of the object. In contrast, our model treats the clustering of the regression coefficients separately from the clustering of the precisions. Thus, two objects having similar observed expression patterns but very different sample variances would usually be placed in separate clusters by BEMMA. Our method, however, would typically cluster their two regression coefficients, but place their precisions in different clusters. Thus the clustering of regression coefficients is decoupled from the issue of heteroscedasticity in our model. Assuming our model more closely reflects the real-world biological process, we would expect the greater flexibility of our SIMTAC method to yield more accuracy in inference than the BEMMA approach.

2.4 Integrating away the means

Notice that the object-specific means $\mu_1, \ldots, \mu_G$ are not used in defining clusters among the regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ nor the precisions $\lambda_1, \ldots, \lambda_G$. Further, the means are not part of hypothesis testing, which involves the regression coefficients. Thus, the means $\mu_1, \ldots, \mu_G$ are nuisance parameters. Dahl and Newton (2007) dealt with them by designating a reference treatment and subtracting the observed data from the mean of the reference treatment. This was an apt approach, leading to a computationally-efficient conjugate DPM model. Unfortunately, their method is not invariant to the choice of the reference treatment and data from the reference treatment is not utilized further.

Because we cluster the regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ and precisions $\lambda_1, \ldots, \lambda_G$ separately, the differencing technique of Dahl and Newton (2007) will not make our model conjugate, and hence the technique is not useful here. We take a more conventional approach to deal with the nuisance parameters; simply integrate the likelihood with respect to the prior distribution of $\mu_1, \ldots, \mu_G$. This has the benefits of not requiring the specification of an arbitrary reference treatment and not losing the data from that treatment. Appendix B shows that, after performing the integration, we are left with the following integrated likelihood that is free of $\mu_1, \ldots, \mu_G$:

$$\boldsymbol{d}_g \mid \boldsymbol{\beta}_g, \lambda_g \sim \mathrm{N}_K\left(\boldsymbol{d}_g \mid \mathrm{X}\boldsymbol{\beta}_g + \mathrm{E}_g^{-1}\boldsymbol{f}_g, \frac{\mathrm{E}_g}{\lambda_g \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}\right), \tag{2.2}$$

where

$$\begin{aligned} \mathbf{E}_g &= \lambda_g(\lambda_g \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)\mathbf{M} - \lambda_g^2 \mathbf{M}\boldsymbol{j}\boldsymbol{j}'\mathbf{M} \quad \text{and} \\ \boldsymbol{f}_g &= \lambda_g m_\mu p_\mu \mathbf{M}\boldsymbol{j}. \end{aligned} \tag{2.3}$$

3 Inference

3.1 Sampling from the posterior distribution

Inference is based on the posterior distribution $p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G, \lambda_1, \ldots, \lambda_G \mid \boldsymbol{d}_1, \ldots, \boldsymbol{d}_G)$. As with all DPM models, the posterior is not available in closed-form. Many Markov chain Monte Carlo (MCMC) samplers and other Monte Carlo techniques have been proposed for DPM models. For reviews and comparisons, see Quintana and Newton (2000) and Neal (2000).

Our implementation fits the model using an MCMC scheme that alternates between a sampler for the regression coefficients and a sampler for the precisions. That is, one sampler updates the configuration of the regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ given the data and the current value of the precisions $\lambda_1, \ldots, \lambda_G$. Then, another sampler updates the configuration of the precisions given the data and the current value of the regression coefficients.

The sampler for each parameter type in our implementation is a hybrid sampler that alternates between the Auxiliary Gibbs sampler of Neal (2000) and Sequentially-Allocated Merge-Split (SAMS) sampler of Dahl (2005). This hybrid approach has the advantage of being capable of exploring the state space with both incremental refinements (via one-at-a-time Gibbs-style updates) and dramatic moves (via merging and splitting clusters). Both the Auxiliary Gibbs sampler and the SAMS sampler concentrate on updating the clustering of the parameter.

Given a clustering $\boldsymbol{\pi}_{\boldsymbol{\beta}} = \{S_1, \ldots, S_q\}$ for the regression coefficients and conditioning on the data and precisions, the values of $\boldsymbol{\beta}_{S_1}, \ldots, \boldsymbol{\beta}_{S_q}$ can be updated by any MCMC sampler, including a random walk sampler and the Gibbs sampler. We recommend the Gibbs sampler, since the full conditional of $\boldsymbol{\beta}_S$ for a cluster S is a standard distribution. Specifically, Appendix C shows that:

$$\boldsymbol{\beta}_S \mid \boldsymbol{d}_1, \ldots, \boldsymbol{d}_G, \lambda_1, \ldots, \lambda_G \sim \mathrm{N}_L(\boldsymbol{\beta}_S \mid \mathbf{U}_S^{-1}\boldsymbol{v}_S, \mathbf{U}_S) \tag{3.1}$$

where

$$\mathbf{U}_S = \mathbf{X}' \sum_{g \in S} (\lambda_g \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)^{-1} \mathbf{E}_g \mathbf{X} + \mathbf{P}_{\boldsymbol{\beta}}$$

Box 1 Computational Procedure

```
1. Initialize the regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$:
  (a) Choose an initial clustering. Two obvious choices are having one cluster for all the coefficients or placing each coefficient in a cluster by itself.
  (b) For each initial cluster $S$ of regression coefficients, initialize the value of the common regression coefficient $\boldsymbol{\beta}_S$ by sampling from the centering distribution $G^{\star}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ of the corresponding DP prior.
2. Initialize the precisions $\lambda_1, \ldots, \lambda_G$:
  (a) Choose an initial clustering. Two obvious choices are having one cluster for all the precisions or placing each precision in a cluster by itself.
  (b) For each initial cluster $S$ of precisions, initialize the value of the common precision $\lambda_S$ by sampling from the centering distribution $G^{\star}_{\lambda}(\lambda)$ of the corresponding DP prior.
3. Obtain draws from the posterior distribution by repeating the following:
  (a) Given the data and the current configuration of the precisions $\lambda_1, \ldots, \lambda_G$, perform the following MCMC updates:
    i. Given the values of the coefficient for each cluster, update the clustering configuration of the coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ using:
      A. One iteration of the Auxiliary Gibbs sampler of Neal (2000).
      B. One iteration of the SAMS sampler of Dahl (2005).
    ii. Given the clustering configuration of the coefficients, update the values of the coefficients using the full conditional distribution in (3.1).
  (b) Given the data and the current configuration of the coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$, perform the following MCMC updates:
    i. Given the values of the precision for each cluster, update the clustering configuration of the precisions $\lambda_1, \ldots, \lambda_G$ using:
      A. One iteration of the Auxiliary Gibbs sampler of Neal (2000).
      B. One iteration of the SAMS sampler of Dahl (2005).
    ii. Given the clustering configuration of the precisions, update the values of the precisions using a random walk having normal proposals with variance $a_\lambda/(5b_\lambda)^2$.
```

$$\boldsymbol{v}_S = \mathbf{X}' \sum_{g \in S} (\lambda_g \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)^{-1} \mathbf{E}_g (\boldsymbol{d}_g - \mathbf{E}_g^{-1} \boldsymbol{f}_g) + \mathbf{P}_{\boldsymbol{\beta}} \boldsymbol{m}_{\boldsymbol{\beta}}.$$

In the case of λ_S, its full conditional is not of a known form. Instead of a Gibbs sampler, we simply use a random walk having normal proposals with variance $a_\lambda/(5b_\lambda)^2$. The computational procedure is summarized algorithmically in Box 1.

Using the set partition parameterization, we denote B (approximate) draws from the posterior distribution as:

$$\boldsymbol{\pi}_{\boldsymbol{\beta}}^{(1)}, \boldsymbol{\phi}_{\boldsymbol{\beta}}^{(1)}, \boldsymbol{\pi}_{\lambda}^{(1)}, \boldsymbol{\phi}_{\lambda}^{(1)}, \ldots, \boldsymbol{\pi}_{\boldsymbol{\beta}}^{(B)}, \boldsymbol{\phi}_{\boldsymbol{\beta}}^{(B)}, \boldsymbol{\pi}_{\lambda}^{(B)}, \boldsymbol{\phi}_{\lambda}^{(B)}, \tag{3.2}$$

where $\boldsymbol{\pi}_{\boldsymbol{\beta}}^{(i)}$ and $\boldsymbol{\pi}_{\lambda}^{(i)}$ are the ith set partitions for the regression coefficients and the precisions, respectively, and $\boldsymbol{\phi}_{\boldsymbol{\beta}}^{(i)}$ and $\boldsymbol{\phi}_{\lambda}^{(i)}$ are the ith vectors of unique regression coefficients and precisions, respectively.

3.2 Multiple hypothesis testing

We make hypothesis-testing inference about the regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ through their posterior distribution. This entails defining a univariate parameter q_g that is appropriate for the particular experimental design $\mathbf{X}$ and then estimating it under squared-error loss using the MCMC output. The objects are then ranked by the expected value of these parameters with respect to their marginal posterior distribution. Time and budget constraints will typically dictate how many interesting objects are pursued.

For concreteness, consider a time-course microarray experiment with two treatments (A and B), three time points, and three replicates at each of the six combinations of treatment and time points. As is typical of time-course microarray experiments, the observations are independent. Hence, $\mathbf{M}$ is the identity matrix. Without loss of generality, let the explicit gene-specific mean μ_g correspond to treatment A at the first time point. Thus, the design matrix $\mathbf{X}$ is 18×5 and we let $\boldsymbol{\beta}_g$ be a vector whose five elements $\beta_{g,1}, \ldots, \beta_{g,5}$ respectively correspond to treatment A at the second time point, treatment A at the third time point, treatment B at the first time points, etc. When interested in differential expression anywhere among the six treatments (that is, the usual global F-test hypothesis in one-way ANOVA), the differential expression parameter might be:

$$q_g = \sum_{i=1}^{5} \beta_{g,i}^2.$$

If instead, we are interested in expression changes between the two groups within a time point and not necessarily across time points, a better choice for the differential expression parameter is:

$$q_g = (\beta_{g,3} - 0)^2 + (\beta_{g,4} - \beta_{g,1})^2 + (\beta_{g,5} - \beta_{g,2})^2. \tag{3.3}$$

3.3 Clustering

We view the clustering of the precisions as merely a device to accommodate heteroscedasticity, yet still permit the sharing of information. In this section, we focus on inference on the clustering of the regression coefficients, but the methods apply equally well to the clustering of the precisions.

The sampling algorithm described in Section 3.1 produces clusterings $\boldsymbol{\pi}_{\boldsymbol{\beta}}^{(1)}, \ldots, \boldsymbol{\pi}_{\boldsymbol{\beta}}^{(B)}$ from the posterior clustering distribution of the regression coefficients. Several methods have been proposed to arrive at a point estimate of the clustering using draws from a posterior clustering distribution. Perhaps the simplest is to select the clustering among those in the MCMC output that maximizes the posterior clustering probability mass function. This maximum *a posteriori* (MAP) clustering corresponds to minimizing the posterior expected loss based on a simple 0-1 loss function. Lau and Green (2006) recently proposed a heuristic to approximate the minimization of a posterior expected loss of the more appealing loss function suggested by Binder (1978) for clustering problems.

For each clustering $\boldsymbol{\pi}_{\boldsymbol{\beta}}$ in $\boldsymbol{\pi}_{\boldsymbol{\beta}}^{(1)}, \ldots, \boldsymbol{\pi}_{\boldsymbol{\beta}}^{(B)}$, an association matrix $\delta(\boldsymbol{\pi}_{\boldsymbol{\beta}})$ of dimension $G \times G$ can be formed whose (i, j) element is $\delta_{i,j}(\boldsymbol{\pi}_{\boldsymbol{\beta}})$, an indicator of whether $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$. Element-wise averaging of these association matrices yields a matrix of estimates $\hat{p}_{i,j}$ of the pairwise probabilities that objects are clustered. Medvedovic and Sivaganesan (2002) and Medvedovic *et al.* (2004) use this pairwise probability matrix as a distance matrix in hierarchical agglomerative clustering.

Dahl (2006) introduced the least-squares clustering estimator which selects the observed clustering that minimizes the sum of squared deviations of its association matrix $\delta(\boldsymbol{\pi}_{\boldsymbol{\beta}})$ from the pairwise probability matrix:

$$\boldsymbol{\pi}_{\boldsymbol{\beta}}^{\mathrm{LS}} = \underset{\boldsymbol{\pi}_{\boldsymbol{\beta}} \in \{\boldsymbol{\pi}_{\boldsymbol{\beta}}^{(1)}, \ldots, \boldsymbol{\pi}_{\boldsymbol{\beta}}^{(B)}\}}{\arg\min} \sum_{i=1}^{G} \sum_{j=1}^{G} (\delta_{i,j}(\boldsymbol{\pi}_{\boldsymbol{\beta}}) - \hat{p}_{i,j})^2. \tag{3.4}$$

Here we use the least-squares clustering method of Dahl (2006). Like the MAP clustering, the least-squares clustering is selected among the clusterings sampled by the Markov chain. The least-squares clustering is the sampled clustering that minimizes the posterior expected loss of Binder (1978) — in this case, assuming equal costs of clustering mistakes. The least-squares clustering is trivial to implement and is not computationally demanding for even large G.

Clustering inference can be more than providing a point estimate of clustering. For example, if two objects are estimated to belong to the same cluster, the pairwise probability matrix provides an estimate of the probability they are not clustered. More generally, the pairwise probability matrix can yield information about the strength of a cluster. Of course, using the MCMC output, it is also trivial to estimate the distribution of the number of clusters and the distribution of the cluster sizes.

4 Simulation study

In order to illustrate the proposed SIMTAC method and investigate its potential benefits with respect to some existing methods, we present a simulation study. We compare our method to ANOVA and the BEMMA method of Dahl and Newton (2007). We use standard software for ANOVA method. Software for SIMTAC and BEMMA are available at `http://www.stat.tamu.edu/~dahl/software`.

4.1 Synthetic data

We used the experimental design described in Section 3.2 that imitates a time-course microarray experiment setting with two treatments, three time points and three replicates at each of the six combinations of treatment and time points. We simulated 50 independent datasets containing 144 differentially expressed genes among 720 genes. Each dataset contained 216 clusters of various sizes, as indicated in Table 1. By definition, genes within a cluster have a common value for their regression coefficients. A gene g within a given cluster has relationships among its regression coefficients $\beta_{g,1}, \ldots, \beta_{g,5}$ as shown in Table 1. Recall that the regression coefficients encode whether the genes in a cluster are differentially or equivalently expressed. Note that some clusters have the same size and relationship among the regression coefficients. In all cases, however, the regression coefficients of the clusters are independent standard normal deviates, subject to the equality constraints shown in Table 1.

The precisions $\lambda_1, \ldots, \lambda_G$ in the synthetic datasets had a much simpler clustering design: Each gene was placed in one of 12 clusters (each of 60 genes) whose precision was independently drawn from the gamma distribution $\text{Ga}(\lambda \mid 10, 10)$ having mean 1. Finally, for each dataset, the clustering of the genes in terms of their precision was randomized with respect to their clustering in terms of the regression coefficients.

4.2 Multiple hypothesis testing

We applied our proposed method to the synthetic datasets using the hyperparameter values recommended in Appendix A. For each dataset, two Markov chains were run from randomly chosen starting states. On standard computers purchased in 2003, each chain was run for 5,000 iterations (except two chains whose allocated two hours of CPU time expired after only 4,890 iterations). Only 1-in-10 iterations were recorded. Trace plots indicated that discarding some iterations for a burn-in was probably not necessary and results from the two chains were combined. We assumed we were interested in expression changes between the two groups within a time point and not necessarily across time points, and thus ranked genes for differential expression using the parameter in (3.3).

Table 1 Clusters in a synthetic dataset

Size of each cluster	Relationship of regression coefficients encoding equivalent and differential expression			Number of clusters with this configuration
	Time point 1	Time point 2	Time point 3	
120	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	1
40	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	2
40	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	1
15	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	6
15	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	1
15	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	1
5	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	19
5	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	2
5	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	2
5	$\beta_{g,3} \neq 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	1
2	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	48
2	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	4
2	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	4
2	$\beta_{g,3} \neq 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	4
1	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	95
1	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	5
1	$\beta_{g,3} = 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	5
1	$\beta_{g,3} \neq 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	5
1	$\beta_{g,3} = 0$	$\beta_{g,4} = \beta_{g,1}$	$\beta_{g,5} \neq \beta_{g,2}$	5
1	$\beta_{g,3} \neq 0$	$\beta_{g,4} \neq \beta_{g,1}$	$\beta_{g,5} = \beta_{g,2}$	5

Note: For the 216 clusters in each synthetic dataset, this table shows the relationship among and the cluster sizes for the regression coefficients. Although some clusters have the same size and relationship among the regression coefficients, in all cases the regression coefficients are independent standard normal deviates (subject to the constraints presented in the table)

For comparison purposes, two other methods for detecting differential gene expression were applied to the synthetic data: BEMMA of Dahl and Newton (2007) and Analysis of Variance (ANOVA). For BEMMA, we used the hyperparameters, burn-in procedure and MCMC samplers recommend by Dahl and Newton (2007). For each dataset, one chain was run from a burned-in state for four hours. We used the same parameter in (3.3) for the BEMMA method. For ANOVA, we performed a full and reduced model test, where the full model had unconstrained regression coefficients and the reduced model constrained the regression coefficients within a time point to be equal. For the ANOVA procedure, we ranked genes for differential expression by their p-values.

The proportion of false discoveries was used to compare the three methods. For each of the 50 independent datasets, the methods provided rankings of the genes in terms of their perception of evidence for differential expression. These lists were truncated at $1, 2, \ldots, 100$ genes. At each truncation, the proportions of false discoveries were computed and averaged over the 50 datasets. Figure 1 shows, for

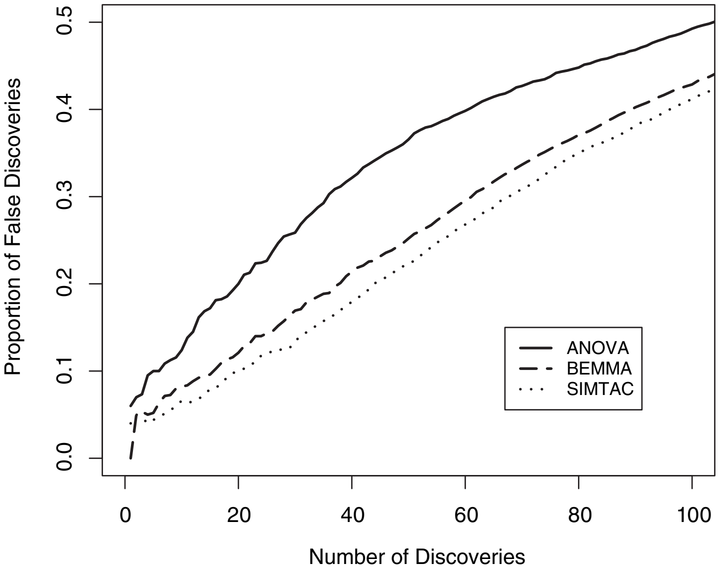

Figure 1 Proportion of false discoveries for ANOVA, BEMMA, SIMTAC methods in simulation study

each method, the estimated proportion of false discoveries as a function of the number of discoveries. Notice that BEMMA and SIMTAC are both substantially better than ANOVA and our method appears to be somewhat better than BEMMA. Figure 2 shows the estimated difference in the proportion of false discoveries from ANOVA and SIMTAC as well as the difference for BEMMA and SIMTAC. After about 10 to 20 discoveries, the differences become statistically significant.

4.3 Clustering

For each dataset, we applied the least-squares clustering technique to the clusterings produced by BEMMA and our proposed method. (Software for the least-squares clustering is available at `http://www.stat.tamu.edu/~dahl/software`) Since the true clustering is known in our simulation study, we could compute the agreement of BEMMA's least-squares clustering with the true clustering. Likewise, we computed the agreement of the least-squares clustering from our method with the true clustering.

There are many procedures for measuring the agreement between two clusterings. In a comprehensive comparison, Milligan and Cooper (1986) recommend the adjusted Rand index (Rand 1971; Hubert and Arabie 1985) as the preferred measure of agreement between two clusterings. Large values for the adjusted Rand index

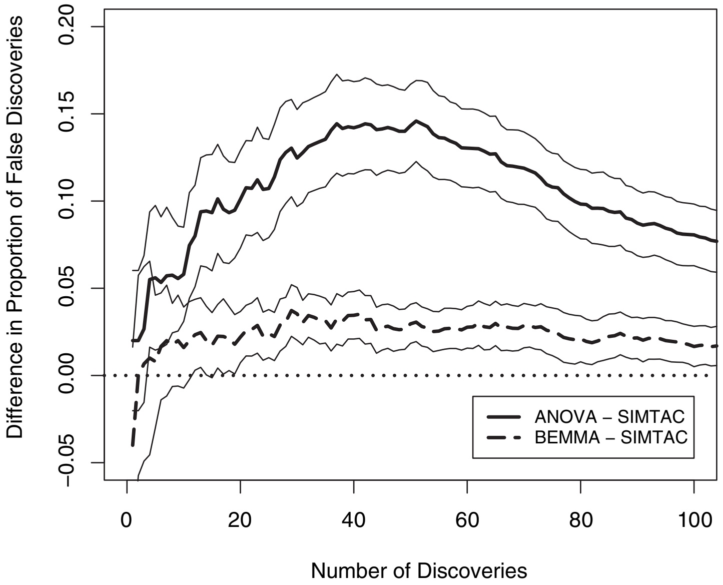

Figure 2 Difference in proportion of false discoveries for ANOVA, BEMMA, SIMTAC methods in simulation study. Plots of the average difference in the proportion of false discoveries between ANOVA and SIMTAC (solid line) and BEMMA and SIMTAC (long, dashed line). Points above the short, dashed reference line at zero indicate worse performance in relation to the proposed SIMTAC method. The thin lines represent 95% pointwise confidence intervals

mean better agreement. That is, an estimated clustering that closely matches the true clustering has a relatively large adjusted Rand index.

We found that in only 18 of the 50 simulated datasets, the least-squares clustering for the regression coefficients in our method had a better adjusted Rand index than that of BEMMA. This suggests that BEMMA better estimates the true clustering of the regression coefficients (two-sided p-value = 0.05). In 45 of the 50 simulated datasets, however, the least-squares clustering for the precisions in our method had a larger adjusted Rand index than that of BEMMA. This result is highly statistically significant.

5 Discussion

The simulation study shows that substantial gains are possible when simultaneously modelling the clustering and parameters in hypothesis testing (for example, using BEMMA or our SIMTAC method) instead of ignoring the dependences among the data (as does ANOVA). We have also demonstrated that our refinements to the original BEMMA method of Dahl and Newton (2007) can yield appreciable improvements. In addition, our method is more widely applicable.

The clustering results indicate that BEMMA and our SIMTAC method seem to be comparable in their accuracy for clustering the regression coefficients, but the SIMTAC method is superior in terms of the precisions. This suggests that BEMMA is forced to compromise the clustering of the precisions in order to accommodate the clustering of the regression coefficients. The fact that our method clusters the regression coefficients separately from the precision seems to be advantageous. This may explain why our SIMTAC method is superior in terms of finding differentially expressed genes in the simulation study.

Finally, our SIMTAC method provides two clusterings: one for the regression coefficients and one for the precisions. The two clusterings are capturing different features of the data and will not necessarily result in similar clusterings. We feel that researchers, on a practical level, would be more interested in the clustering of the regression coefficients. Indeed, rather than interpreting the clusters for the precisions, our view is that they are merely a device to accommodate heteroscedasticity yet still permit the sharing of information.

Acknowledgements

The authors thank Fernando Quintana, Peter Müller and an anonymous referee for very helpful comments that improved the manuscript. They also thank Jaesik Jeong for help with some initial calculations. Marina Vannucci is partially supported by NIH/NHGRI grant R01HG003319 and by NSF award DMS 0605001. All computations were performed using the SIMTAC software, available at http://www.stat.tamu.edu/∼dahl/software/simtac.

References

Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, **2**, 1152–174.

Binder DA (1978) Bayesian cluster analysis. *Biometrika*, **65**, 31–38.

Dahl DB (2005) Sequentially-allocated merge-split sampler for Dirichlet process mixture models. Submitted to *Journal of Computational and Graphical Statistics*.

Dahl DB (2006) Model-based clustering for expression data via a Dirichlet process mixture model. In KA DO, P Müller and M Vannucci eds, *Bayesian inference for gene expression and proteomics*, pp. 201–218. USA: Cambridge University Press.

Dahl DB and Newton MA (2007) Multiple hypothesis testing by clustering treatment effects. *Journal of the American Statistical Association*, **102**, 517–26.

Escobar MD and West M (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–88.

Ferguson, TS (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–30.

Hubert L and Arabie P (1985) Comparing partitions. *Journal of Classification*, **2**, 193–218.

Lau JW and Green PJ (2006) Bayesian model based clustering procedures. *Technical report*. Department of Mathematics, University of Bristol.

Medvedovic M, and Sivaganesan S (2002) Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatrics*, **18**, 1194–206.

Medvedovic M, Yeung K and Bumgarner R (2004) Bayesian mixture model based clustering of replicated microarray data. *Bioinformatrics*, **20**, 1222–232.

Milligan GW and Cooper MC (1986) A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, **21**, 441–58.

Müller P and Quintana F (2004) Nonparametric Bayesian data analysis. *Statistical Science*, **19**, 95–110.

Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–65.

Quintana FA and Newton MA (2000) Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences. *Journal of Computational and Graphical Statistics*, **9**, 711–37.

Rand WM (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–50.

Tibshirani R and Wasserman L (2006) Correlation-sharing for detection of differential gene expression. *Technical report* 839, Department of Statistics, Carnegie Mellon University.

Yuan M and Kendziorski C (2006) A unified approach for simultaneous gene clustering and differential expression identification. *Biometrics*, **62**, 1089–98.

Appendix A: Setting the hyperparameters

Our recommendation for setting the hyperparameters is based on computing for each object the least-squares estimates of the regression coefficients, y-intercept and mean-squared error. We set m_μ to be the mean of the estimated y intercepts and p_μ to be the inverse of their variances. We set $\mathbf{P}_{\boldsymbol{\beta}}$ to be an identity matrix times the average precision of the estimated regression coefficients within a gene. Finally, a_λ and b_λ are set using method of moments estimation, assuming that the inverse of the mean-squared errors are random draws from a gamma distribution having mean a_λ / b_λ.

Appendix B: Integrated likelihood

Dropping the subscript g for simplicity and integrating with respect to the nuisances parameters $\mu_1, \ldots, \mu_g$, the marginal likelihood is:

$$p\left(\boldsymbol{d} \mid \boldsymbol{\beta}, \lambda\right) = \int p\left(\boldsymbol{d} \mid \boldsymbol{\beta}, \lambda, \mu\right) p\left(\mu\right) d\mu$$

$$\propto \int \exp\left\{ -\frac{1}{2}\left[(\boldsymbol{d} - \mathbf{X}\boldsymbol{\beta} - \mu \boldsymbol{j})' \lambda \mathbf{M} (\boldsymbol{d} - \mathbf{X}\boldsymbol{\beta} - \mu\boldsymbol{j}) + p_\mu \left(\mu - m_\mu\right)^2 \right]\right\} d\mu$$

Letting $\boldsymbol{a} = \boldsymbol{d} - \mathbf{X}\boldsymbol{\beta}$ and working with most of the exponent above, we have:

$$\begin{aligned}
&(\boldsymbol{d} - \mathbf{X}\boldsymbol{\beta} - \mu \boldsymbol{j})' \lambda \mathbf{M} (\boldsymbol{d} - \mathbf{X}\boldsymbol{\beta} - \mu\boldsymbol{j}) + p_\mu \left(\mu - m_\mu\right)^2 \\
&\quad = \lambda \left(\boldsymbol{a} - \mu \boldsymbol{j}\right)' \mathbf{M} \left(\boldsymbol{a} - \mu \boldsymbol{j}\right) + p_\mu \left(\mu - m_\mu\right)^2 \\
&\quad = \lambda \boldsymbol{a}'\mathbf{M}\boldsymbol{a} - 2\lambda(\mu\boldsymbol{j})'\mathbf{M}\boldsymbol{a} + \lambda(\mu\boldsymbol{j})'\mathbf{M}\mu\boldsymbol{j} + p_\mu(\mu^2 - 2\mu m_\mu + m_\mu^2) \\
&\quad = (\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)\mu^2 - 2(\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{a} + m_\mu p_\mu)\mu + \boldsymbol{a}'\lambda\mathbf{M}\boldsymbol{a} + p_\mu m_\mu^2 \\
&\quad = (\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)\left(\mu^2 - 2\left(\frac{\boldsymbol{j}'\lambda\mathbf{M}\boldsymbol{a} + m_\mu p_\mu}{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}\right)\mu + \left(\frac{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{a} + m_\mu p_\mu}{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}\right)^2\right) \\
&\qquad + \lambda\boldsymbol{a}'\mathbf{M}\boldsymbol{a} + p_\mu m_\mu^2 - \frac{(\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{a} + m_\mu p_\mu)^2}{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu} \\
&\quad = (\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)\left(\mu - \frac{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{a} + m_\mu p_\mu}{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}\right)^2 + \lambda\boldsymbol{a}'\mathbf{M}\boldsymbol{a} + p_\mu m_\mu^2 - \frac{(\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{a} + m_\mu p_\mu)^2}{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}
\end{aligned}$$

Therefore:

$$\begin{aligned}
p\left(\boldsymbol{d} \mid \boldsymbol{\beta}, \lambda\right) &\propto \exp\left\{ -\frac{1}{2}\left(\lambda\boldsymbol{a}'\mathbf{M}\boldsymbol{a} + p_\mu m_\mu^2 - \frac{(\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{a} + m_\mu p_\mu)^2}{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}\right)\right\} \\
&\quad \times \int \frac{(\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)\left(\mu - \frac{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{a} + m_\mu p_\mu}{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}\right)^2\right\} d\mu \\
&= \exp\left\{ -\frac{1}{2}\left(\lambda\boldsymbol{a}'\mathbf{M}\boldsymbol{a} - \frac{\lambda^2\boldsymbol{a}'\mathbf{M}\boldsymbol{j}\,\boldsymbol{j}'\mathbf{M}\boldsymbol{a} + 2\lambda m_\mu p_\mu \boldsymbol{j}'\mathbf{M}\boldsymbol{a} + m_\mu^2 p_\mu^2}{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}\right)\right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2}\left(\frac{\lambda(\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)\boldsymbol{a}'\mathbf{M}\boldsymbol{a} - \lambda^2\boldsymbol{a}'\mathbf{M}\boldsymbol{j}\,\boldsymbol{j}'\mathbf{M}\boldsymbol{a} - 2\lambda m_\mu p_\mu \boldsymbol{j}'\mathbf{M}\boldsymbol{a}}{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}\right)\right\} \\
&= \exp\left\{ -\frac{1}{2}\left(\frac{\boldsymbol{a}'(\lambda(\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)\mathbf{M} - \lambda^2\mathbf{M}\boldsymbol{j}\,\boldsymbol{j}'\mathbf{M})\boldsymbol{a} - 2\lambda m_\mu p_\mu \boldsymbol{j}'\mathbf{M}\boldsymbol{a}}{\lambda\boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}\right)\right\}
\end{aligned}$$

For simplicity, define the following:

$$\begin{aligned}\mathrm{E} &= \lambda(\lambda \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)\mathbf{M} - \lambda^2 \mathbf{M}\boldsymbol{j}\,\boldsymbol{j}'\mathbf{M}\\ \boldsymbol{f} &= \lambda m_\mu p_\mu \mathbf{M}\boldsymbol{j}\end{aligned}$$

Then:

$$\begin{aligned}p\,(\boldsymbol{d} \mid \boldsymbol{\beta}, \lambda) &\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{a} - \mathrm{E}^{-1}\boldsymbol{f})'\frac{\mathrm{E}}{\lambda \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}(\boldsymbol{a} - \mathrm{E}^{-1}\boldsymbol{f})\right\}\\ &= \exp\left\{ -\frac{1}{2}(\boldsymbol{d} - \mathrm{X}\boldsymbol{\beta} - \mathrm{E}^{-1}\boldsymbol{f})'\frac{\mathrm{E}}{\lambda \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}(\boldsymbol{d} - \mathrm{X}\boldsymbol{\beta} - \mathrm{E}^{-1}\boldsymbol{f})\right\},\end{aligned}$$

which is the kernel of the multivariate normal distribution given in (2.2).

Appendix C: Full conditional of regression coefficients

Below we give the full conditional of the regression coefficient for a cluster S. For convenience, let $\mathrm{Q}_g = (\lambda_g \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)^{-1}\mathrm{E}_g$, $\mathrm{E}_g = \lambda_g(\lambda_g \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu)\mathbf{M} - \lambda_g^2 \mathbf{M}\boldsymbol{j}\,\boldsymbol{j}'\mathbf{M}$ and $\boldsymbol{f}_g = \lambda_g m_\mu p_\mu \mathbf{M}\boldsymbol{j}$. Also, let $\boldsymbol{d}_S$ and λ_S be the collection of observations and precisions corresponding to the cluster S. Note that, whereas $\boldsymbol{\beta}_S$ is a single value shared by all the genes in cluster S, the values of the precisions in λ_S will likely not be constant within S, since S is not a cluster of the precisions. The full conditional is:

$$\begin{aligned}&p(\boldsymbol{\beta}_S \mid \boldsymbol{d}_S, \lambda_S) \propto p(\boldsymbol{d}_S \mid \boldsymbol{\beta}_S, \lambda_S)p(\boldsymbol{\beta}_S)\\ &\quad\propto \exp\left\{ -\tfrac{1}{2}\left(\sum_{g\in S}(\boldsymbol{d}_g - \mathrm{X}\boldsymbol{\beta}_S - \mathrm{E}_g^{-1}\boldsymbol{f}_g)'\mathrm{Q}_g(\boldsymbol{d}_g - \mathrm{X}\boldsymbol{\beta}_S - \mathrm{E}_g^{-1}\boldsymbol{f}_g)\right.\right.\\ &\qquad\left.\left. + (\boldsymbol{\beta}_S - \boldsymbol{m}_{\boldsymbol{\beta}})'\mathrm{P}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_S - \boldsymbol{m}_{\boldsymbol{\beta}})\right)\right\}\\ &\quad\propto \exp\left\{ -\tfrac{1}{2}\left(\sum_{g\in S}\left(\boldsymbol{\beta}_S'\mathrm{X}'\mathrm{Q}_g\mathrm{X}\boldsymbol{\beta}_S - 2\boldsymbol{\beta}_S'\mathrm{X}'\mathrm{Q}_g(\boldsymbol{d}_g - \mathrm{E}_g^{-1}\boldsymbol{f}_g)\right)\right.\right.\\ &\qquad\left.\left. + \boldsymbol{\beta}_S'\mathrm{P}_{\boldsymbol{\beta}}\boldsymbol{\beta}_S - 2\boldsymbol{\beta}_S'\mathrm{P}_{\boldsymbol{\beta}}\boldsymbol{m}_{\boldsymbol{\beta}}\right)\right\}\\ &\quad\propto \exp\left\{ -\tfrac{1}{2}\left(\boldsymbol{\beta}_S'(\mathrm{X}'\sum_{g\in S}\mathrm{Q}_g\mathrm{X} + \mathrm{P}_{\boldsymbol{\beta}})\boldsymbol{\beta}_S - 2\boldsymbol{\beta}_S'(\mathrm{X}'\textstyle\sum_{g\in S}\mathrm{Q}_g(\boldsymbol{d}_g - \mathrm{E}_g^{-1}\boldsymbol{f}_g) + \mathrm{P}_{\boldsymbol{\beta}}\boldsymbol{m}_{\boldsymbol{\beta}})\right)\right\}\\ &\quad= \exp\left\{ -\frac{1}{2}(\boldsymbol{\beta}_S - \mathrm{U}_S^{-1}\boldsymbol{v}_S)'\mathrm{U}_S(\boldsymbol{\beta}_S - \mathrm{U}_S^{-1}\boldsymbol{v}_S)\right\}\end{aligned}$$

which is the kernel of the multivariate normal distribution given in (3.1).