Published in final edited form as:
Bayesian Anal. ; 4(4): 707–732. doi:10.1214/09-BA426.

Spiked Dirichlet Process Prior for Bayesian Multiple Hypothesis Testing in Random Effects Models

Sinae Kim[*], **David B. Dahl**[†], and **Marina Vannucci**[‡]

[*]Department of Biostatistics, University of Michigan, Ann Arbor, MI, sinae@umich.edu

[†]Department of Statistics, Texas A&M University, College Station, TX, dahl@stat.tamu.edu

[‡]Department of Statistics, Rice University, Houston, TX, marina@rice.edu

Abstract

We propose a Bayesian method for multiple hypothesis testing in random effects models that uses Dirichlet process (DP) priors for a nonparametric treatment of the random effects distribution. We consider a general model formulation which accommodates a variety of multiple treatment conditions. A key feature of our method is the use of a product of spiked distributions, i.e., mixtures of a point-mass and continuous distributions, as the centering distribution for the DP prior. Adopting these spiked centering priors readily accommodates sharp null hypotheses and allows for the estimation of the posterior probabilities of such hypotheses. Dirichlet process mixture models naturally borrow information across objects through model-based clustering while inference on single hypotheses averages over clustering uncertainty. We demonstrate via a simulation study that our method yields increased sensitivity in multiple hypothesis testing and produces a lower proportion of false discoveries than other competitive methods. While our modeling framework is general, here we present an application in the context of gene expression from microarray experiments. In our application, the modeling framework allows simultaneous inference on the parameters governing differential expression and inference on the clustering of genes. We use experimental data on the transcriptional response to oxidative stress in mouse heart muscle and compare the results from our procedure with existing nonparametric Bayesian methods that provide only a ranking of the genes by their evidence for differential expression.

Keywords

Bayesian nonparametrics; differential gene expression; Dirichlet process prior; DNA microarray; mixture priors; model-based clustering; multiple hypothesis testing

1 Introduction

This paper presents a semiparametric Bayesian approach to multiple hypothesis testing in random effects models. The model formulation borrows strength across similar objects (here, genes) and provides probabilities of sharp hypotheses regarding each object.

Much of the literature in multiple hypothesis testing has been driven by DNA microarrays studies, where gene expression of tens of thousands of genes are measured simultaneously (Dudoit et al. 2003). Multiple testing procedures seek to ensure that the family-wise error rate (FWER) (e.g., Hochberg (1988), Hommel (1988), Westfall and Young (1993)), the false discovery rate (FDR) (e.g., Benjamini and Hochberg (1995), Storey (2002), Storey (2003), and Storey et al. (2004)), or similar quantities (e.g., Newton et al. (2004)) are below a nominal level without greatly sacrificing power. Accounts on the Bayesian perspective to multiple testing are provided by Berry and Hochberg (1999) and Scott and Berger (2006).

There is a great variety of modeling settings that accommodate multiple testing procedures. The simplest approach, extensively used in the early literature on microarray data analysis, is to apply standard statistical procedures (such as the *t*-test) separately and then combine the results for simultaneous inference (e.g., Dudoit et al. 2002). Westfall and Wolfinger (1997) recommended procedures that incorporate dependence. Baldi and Long (2001), Newton et al. (2001), Do et al. (2005) and others have sought prior models that share information across objects, particularly when estimating object-specific variance across samples. Yuan and Kendziorski (2006) use finite mixture models to model dependence. Classical approaches that have incorporated dependence in the analysis of gene expression data include Tibshirani and Wasserman (2006), Storey et al. (2007), and Storey (2007) who use information from related genes when testing for differential expression of individual genes.

Nonparametric Bayesian approaches to multiple testing have also been explored (see, for example, Gopalan and Berry (1998), Dahl and Newton (2007), MacLehose et al. (2007), Dahl et al. (2008)). These approaches model the uncertainty about the distribution of the parameters of interest using Dirichlet process (DP) prior models that naturally incorporate dependence in the model by inducing clustering of similar objects. In this formulation, inference on single hypotheses is typically done by averaging over clustering uncertainty. Dahl and Newton (2007) and Dahl et al. (2008) show that this approach leads to increased power for hypothesis testing. However, the methods provide posterior distributions that are continuous, and cannot therefore be used to directly test sharp hypotheses, which have zero posterior probability. Instead, decisions regarding such hypotheses are made based on calculating univariate scores that are context specific. Examples include the sum-of-squares of the treatment effects (to test a global ANOVA-like hypothesis) and the probability that a linear combination of treatment effects exceeds a threshold.

In this paper we build on the framework of Dahl and Newton (2007) and Dahl et al. (2008) to show how the DP modeling framework can be adapted to provide meaningful posterior probabilities of sharp hypotheses by using a mixture of a point-mass and a continuous distribution as the centering distribution of the DP prior on the coefficients of a random effects model. This modification retains the increased power of DP models but also readily accommodates sharp hypotheses. The resulting posterior probabilities have a very natural interpretation in a variety of uses. For example, they can be used to rank objects and define a list according to a specified expected number of false discoveries. We demonstrate via a simulation study that our method yields increased sensitivity in multiple hypothesis testing and produces a lower proportion of false discoveries than other competitive methods, including standard ANOVA procedures. In our application, the modeling framework we adopt simultaneously infers the parameters governing differential expression and clusters the objects (i.e., genes). We use experimental data on the transcriptional response to oxidative stress in mouse heart muscle and compare results from our procedure with that of existing nonparametric Bayesian methods which only provide a ranking of the genes by their evidence for differential expression.

Recently Cai and Dunson (2007) independently proposed the use of similar spiked priors in DP priors in a Bayesian nonparametric linear mixed model where variable selection is achieved by modeling the unknown distribution of univariate regression coefficients. Similarly, MacLehose et al. (2007) used this formulation in their DP mixture model to account for highly correlated regressors in an observational study. There, the clustering induced by the Dirichlet process is on the *univariate* regression coefficients and strength is borrowed *across covariates*. Finally, Dunson et al. (2008) use a similar spiked centering distribution of *univariate* regression coefficients in a logistic regression. In contrast, our goal is nonparametric modeling of multivariate random effects which may equal the zero vector. That is, we do *not* share information across univariate covariates but rather seek to leverage

similarities *across genes* by clustering *vectors* of regression coefficients associated with the genes.

The remainder of the paper is organized as follows. Section 2 describes our proposed modeling framework and the prior model. In Section 3 we discuss the MCMC algorithm for inference. Using simulated data, we show in Section 4.1 how to make use of the posterior probabilities of hypotheses of interest to aid the interpretation of the hypothesis testing results. Section 4.2 describes the application to DNA microarrays. In both Sections 4.1 and 4.2, we compare our proposed method to the LIMMA (Smyth 2004), to the SIMTAC method of Dahl et al. (2008) and to a standard ANOVA procedure. Section 5 concludes the paper.

2 Dirichlet Process Mixture Models for Multiple Testing

2.1 Random Effects Model

Suppose there are K observations on each of G objects and T^* treatments. For each object g, with $g = 1, \ldots, G$, we model the data vector $\boldsymbol{d}_g$ with the following K-dimensional multivariate normal distribution:

$$\boldsymbol{d}_g|\mu_g, \beta_g, \lambda_g \sim \mathrm{N}_K\left(\boldsymbol{d}_g|\mu_g \boldsymbol{j} + \mathbf{X}\beta_g, \lambda_g \mathbf{M}\right), \quad (1)$$

where μ_g is an object-specific mean, $\boldsymbol{j}$ is a vector of ones, $\mathbf{X}$ is a $K \times T$ design matrix, β_g is a vector of T regression coefficients specific to object g, $\mathbf{M}$ is the inverse of a correlation matrix of the K observations from an object, and λ_g is an object-specific precision (i.e., inverse of the variance). We are interested in testing a hypothesis for each of G objects in the form:

$$\begin{aligned} &H_{0,g}: \beta_{1,g} = \ldots = \beta_{T^*,g} = 0 \\ &H_{a,g}: \beta_{t,g} \neq 0 \text{ for some } t = 1, \ldots, T^* \end{aligned} \quad (2)$$

for $g = 1, \ldots, G$.

Object-specific intercept terms are $\mu_g \boldsymbol{j}$, so the design matrix $\mathbf{X}$ does not contain the usual column of ones and T is one less than the number of treatments (i.e., $T = T^* - 1$). Also, $d_1, \ldots, d_G$ are assumed to be conditionally independent given all model parameters. In the example of Section 4.2, the objects are genes with $\boldsymbol{d}_g$ being the background-adjusted and normalized expression data for a gene g under T^* treatments, G being the number of genes, and K being the number of microarrays. In the example, we have $K = 12$ since there are 3 replicates for each of $T^* = 4$ treatments, and the $\mathbf{X}$ matrix is therefore:

$$\mathbf{X} = \begin{pmatrix} 0_3 & 0_3 & 0_3 \\ \boldsymbol{j}_3 & 0_3 & 0_3 \\ 0_3 & \boldsymbol{j}_3 & 0_3 \\ 0_3 & 0_3 & \boldsymbol{j}_3 \end{pmatrix}$$

where $\boldsymbol{j}_3$ is a 3-dimensional column vector of ones and 0_3 a 3-dimensional column vector of zeroes. If there are other covariates available, they would be placed as extra columns in $\mathbf{X}$. Note that the design matrix $\mathbf{X}$ and the correlation matrix $\mathbf{M}$ are known and common to all objects, whereas μ_g, β_g, and λ_g are unknown object-specific parameters. For experimental designs involving independent sampling (e.g., the typical time-course microarray experiment in which subjects are sacrificed rather than providing repeated measures), $\mathbf{M}$ is simply the identity matrix.

2.2 Prior Model

We take a nonparametric Bayesian approach to model the uncertainty on the distribution of the random effects. The modeling framework we adopt allows for simultaneous inference on the regression coefficients and on the clustering of the objects (i.e., genes). We achieve this by placing a Dirichlet process (DP) prior (Antoniak 1974) with a spiked centering distribution on the distribution function of the regression coefficient vectors, $\beta_1, \ldots, \beta_G$,

$$\begin{aligned} \beta_1, \ldots, \beta_G | \quad & G_\beta \sim G_\beta \\ & G_\beta \sim DP\left(\alpha_\beta, G_{0\beta}\right) \end{aligned}$$

where G_β denotes a distribution function of $\boldsymbol{\beta}$, *DP* stands for a Dirichlet process, α_β is a precision parameter, and $G_{0\beta}$ is a centering distribution, i.e., $E[G_\beta] = G_{0\beta}$. Sampling from *DP* induces ties among $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$, since there is a positive probability that $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for every $i \neq j$. 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 among 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. The fact that our model induces ties among the regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ is the means by which it borrows strength across objects for estimation.

Set partition notation is helpful throughout the paper. A set partition $\xi = \{S_1, \ldots, S_q\}$ of $S_0 = \{1, 2, \ldots, G\}$ has the following properties: Each component S_i is non-empty, the intersection of two components S_i and S_j is empty, and the union of all components is S_0. A cluster S in the set partition ξ for the 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 coefficient vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ can be reparametrized as a partition $\xi_{\boldsymbol{\beta}}$ and a collection of unique model parameters $\phi_{\boldsymbol{\beta}} = (\boldsymbol{\beta}_{S_1}, \ldots, \boldsymbol{\beta}_{S_q})$. We will use the terms clustering and set partition interchangeably.

Spiked Prior Distribution on the Regression Coefficients—Similar modeling frameworks and inferential goals to the one we describe in this paper were considered by Dahl and Newton (2007) and Dahl et al. (2008). However, their prior formulation does not naturally permit hypothesis testing of sharp hypotheses, i.e., it can not provide $\Pr(H_{a,g}|\text{data}) = 1 - \Pr(H_{0,g}|\text{data})$, where hypotheses are defined as in (2), since the posterior distribution of $\beta_{t,g}$ is continuous. Therefore, they must rely on univariate scores capturing evidence for these hypotheses. The prior formulation we adopt below, instead, allows us to estimate the probability of sharp null hypotheses directly from the MCMC samples.

These distributions have been widely used as prior distribution in the Bayesian variable selection literature (George and McCulloch 1993; Brown et al. 1998). Spiked distributions are a mixture of two distributions: the "spike" refers to a point mass distribution at zero and the other distribution is a continuous distribution for the parameter if it is not zero. Here we employ these priors to perform nonparametric multiple hypothesis testing by specifying a spiked distribution as the centering distribution for the DP prior on the regression coefficient vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$. Adopting a spiked centering distribution in DP allows for a positive posterior probability on $\beta_{t,g} = 0$, so that our proposed model is able to provide probabilities of sharp null hypotheses (e.g., $H_{0,g}: \beta_{1,g} = \ldots = \beta_{T^*,g} = 0$ for $g = 1, \ldots, G$) while simultaneously borrowing strength from objects likely to have the same value of the regression coefficients.

We also adopt a "super-sparsity" prior on the probability of $\beta_{t,g} = 0$ (defined as $\boldsymbol{\pi}_t$ for all g), since it is not uncommon that changes in expressions for many genes will be minimal across treatments. The idea of the "super-sparsity" prior was investigated in Lucas et al. (2006). By using another layer in the prior for $\boldsymbol{\pi}_t$, the probability of $\beta_{t,g} = 0$ will be shrunken toward one for genes showing no changes in expressions across treatment conditions.

Specifically, our model uses the following prior for the regression coefficients $\beta_1, \ldots, \boldsymbol{\beta}_G$

$$\begin{aligned} \beta_1, \ldots, \beta_G | G_\beta &\sim G_\beta \\ G_\beta &\sim DP\left(\alpha_\beta, G_{0\beta}\right) \\ G_{0\beta} &= \prod_{t=1}^{T} \left\{\pi_t \delta_0\left(\beta_{t,g}\right) + (1-\pi_t) N\left(\beta_{t,g} | m_t, \tau_t\right)\right\} \end{aligned}$$

$$\pi_1, \ldots, \pi_T | \rho_1, \ldots, \rho_T \sim \prod_{t=1}^{T} \left\{(1-\rho_t)\, \delta_0(\pi_t) + \rho_t \text{Beta}(\pi | a_\pi, b_\pi)\right\}$$

$$\rho_1, \ldots, \rho_T \sim \text{Beta}\left(\rho | a_\rho, b_\rho\right)$$

$$\tau_1, \ldots, \tau_T \sim \text{Gamma}(\tau | a_\tau, b_\tau)$$

Note that a spiked formulation is used for each element of the regression coefficient vector and $\boldsymbol{\pi}_t = p(\beta_{t,1} = 0) = \ldots = p(\beta_{t,G} = 0)$. Typically, $m_t = 0$, but other values may be desired. We use the parameterization of the gamma distribution where the expected value of $\boldsymbol{\tau}_t$ is $a_\tau b_\tau$. For simplicity, let $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_T)$ and $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_T)$.

After marginalized over $\boldsymbol{\pi}_t$ for all t, the $G_{0\beta}$ becomes

$$G_{0\beta} = \prod_{t=1}^{T} \left\{\rho_t r_\pi \delta_0(\beta_t) + (1 - \rho_t r_\pi) N(\beta_t | m_t, \tau_t)\right\},$$

$$\rho_1, \cdots, \rho_t \sim \text{Beta}\left(\rho | a_\rho, b_\rho\right).$$

where $r_{\boldsymbol{\pi}} = a_{\boldsymbol{\pi}}/(a_{\boldsymbol{\pi}} + b_{\boldsymbol{\pi}})$. As noted in equation above, the $\rho_t r_{\boldsymbol{\pi}}$ is now specified as a probability of $\beta_{t,g} = 0$ for all g.

Prior Distribution on the Precisions—Our model accommodates heteroscedasticity while preserving parsimony by placing a DP prior on the precisions: $\lambda_1, \ldots, \lambda_G$:

$$\begin{aligned} \lambda_1, \ldots \lambda_G | G_\lambda &\sim G_\lambda \\ G_\lambda &\sim DP(\alpha_\lambda, G_{0\lambda}) \\ G_{0\lambda} &= \text{Gamma}(\lambda | a_\lambda, b_\lambda) \end{aligned}$$

Note that the clustering of the regression coefficients is separate from that of the precisions. Although this treatment for the precisions also has the effect of clustering the data, we are typically more interested in the clustering from the regression coefficients since they capture changes across treatment conditions. We let ξ_λ denote the set partition for the precisions $\lambda_1, \ldots, \lambda_G$ and let $\boldsymbol{\phi}_\lambda = (\lambda_{S1}, \ldots, \lambda_{Sq})$ be the collection of unique precision values.

Prior Distribution on the Precision Parameters for DP—Following Escobar and West (1995), we place independent Gamma priors on the precision parameters α_β and α_λ of the DP priors:

$$\alpha_\beta \sim \text{Gamma}\left(\alpha_\beta | a_{\alpha\beta}, b_{\alpha\beta}\right),$$

$$\alpha_\lambda \sim \text{Gamma}\left(\alpha_\lambda | a_{\alpha\lambda}, b_{\alpha\lambda}\right).$$

Prior Distribution on the Means—We assume a Gaussian prior on the object-specific mean parameters $\mu_1, \ldots, \mu_G$:

$$\mu_g \sim N\left(\mu_g | m_\mu, p_\mu\right). \quad (3)$$

3 Inferential Procedures

In this section, we describe how to conduct multiple hypothesis tests and clustering inference in the context of our model. We treat the object-specific means $\mu_1, \ldots, \mu_G$ as nuisance parameters since they are not used either in forming clusters or for multiple testing. Thus, we integrate the likelihood with respect to their prior distribution in (3). Simple calculations lead to the following integrated likelihood (Dahl et al. 2008):

$$\boldsymbol{d}_g | \beta_g, \lambda_g \sim \text{N}_K\left(\boldsymbol{d}_g | \mathbf{X}\beta_g + \mathbf{E}_g^{-1}\boldsymbol{f}_g, \frac{\mathbf{E}_g}{\lambda_g \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu}\right), \quad (4)$$

where

$$\begin{aligned} \mathbf{E}_g &= \lambda_g\left(\lambda_g \boldsymbol{j}'\mathbf{M}\boldsymbol{j} + p_\mu\right)\mathbf{M} - \lambda_g^2 \mathbf{M}\boldsymbol{j}\boldsymbol{j}'\mathbf{M}, \quad \text{and} \\ \boldsymbol{f}_g &= \lambda_g m_\mu p_\mu \mathbf{M}\boldsymbol{j}. \end{aligned} \quad (5)$$

Inference is based on the marginal posterior distribution of the regression coefficients, i.e., $p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G \mid \boldsymbol{d}_1, \ldots, \boldsymbol{d}_G)$ or, equivalently, $p(\xi_{\boldsymbol{\beta}}, \boldsymbol{\phi}_{\boldsymbol{\beta}} \mid \boldsymbol{d}_1, \ldots, \boldsymbol{d}_G)$. This distribution is not available in closed-form, so we use a Markov chain Monte Carlo (MCMC) to sample from the full posterior distribution $p(\xi_{\boldsymbol{\beta}}, \boldsymbol{\phi}_{\boldsymbol{\beta}}, \boldsymbol{\phi}_\lambda, \boldsymbol{\phi}_\lambda, \boldsymbol{\rho}, \boldsymbol{\tau} \mid \boldsymbol{d}_1, \ldots, \boldsymbol{d}_G)$ and marginalize over the parameters ξ_λ, $\boldsymbol{\phi}_\lambda$, $\boldsymbol{\rho}$, and $\boldsymbol{\tau}$.

3.1 MCMC Scheme

Our MCMC sampling scheme updates each of the following parameters, one at a time: $\xi_{\boldsymbol{\beta}}$, $\boldsymbol{\phi}_{\boldsymbol{\beta}}$, ξ_λ, $\boldsymbol{\phi}_\lambda$, $\boldsymbol{\rho}$, and $\boldsymbol{\tau}$. Recall that $\boldsymbol{\beta}_S$ is the element of $\boldsymbol{\phi}_{\boldsymbol{\beta}}$ associated with cluster $S \in \xi_{\boldsymbol{\beta}}$, with β_{St} being element t of that vector. Likewise, λ_S is the element of $\boldsymbol{\phi}_\lambda$ associated with cluster $S \in \xi_\lambda$. Given starting values for these parameters, we propose the following MCMC sampling scheme. Details for the first three updates are available in the Appendix.

(1) Obtain draws $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_T)$ from its full conditional distribution by the following procedure. First, sample $Y_t = r_{\boldsymbol{\pi}} \rho_t$ from its conditional distributions:

$$y_t|\cdot \sim p(y_t)\, y_t^{\sum_{S\in\xi} I(\beta_{St}=0)} (1-y_t)^{\sum_{S\in\xi} I(\beta_{St}\neq 0)},$$

with

$$p(y_t) \propto y_t^{a_\rho + b_\rho - 1} (r_\pi - y_t)^{b_{\rho-1}},$$

which does not have a known distributional form. A grid-based inverse-cdf method has been adopted for sampling y_t. Once we draw samples of Y_t, then we will obtain ρ_t as $Y_t / r_{\boldsymbol{\pi}}$.

(2) Draw samples of $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_T)$ from their full conditional distributions:

$$\tau_t|\cdot \sim \text{Gamma}\left(a_\tau + \frac{|\zeta_t|}{2}, \left(\frac{1}{b_\tau} + \frac{1}{2}\sum_{S\in\zeta_t} (\beta_{St} - m_t)^2\right)^{-1}\right), \quad (6)$$

where $\zeta_t = \{S \in \xi_{\boldsymbol{\beta}} \mid \beta_{St} \neq 0\}$ and $|\zeta_t|$ is its cardinality.

(3) Draw samples of $\boldsymbol{\beta}_S = (\beta_{S1}, \ldots, \beta_{ST})$ for their full conditional distributions:

$$\beta_{St}|\cdot \sim \pi_{St}\delta_0 + (1-\pi_{St})\, N\left(h_t^{-1} z_t, h_t\right), \quad (7)$$

where

$$h_t = \tau_t + \sum_{g\in S} \boldsymbol{x}_t^T \boldsymbol{Q}_g \boldsymbol{x}_t,$$

$$z_t = m_t \tau_t + \sum_{g\in S} \boldsymbol{x}_t^T \boldsymbol{Q}_g \boldsymbol{A}_g,$$

$$\boldsymbol{Q}_g = \left(\lambda_g \boldsymbol{j}' \boldsymbol{M} \boldsymbol{j} + p_\mu\right)^{-1} \boldsymbol{E}_g,$$

$$\boldsymbol{A}_g = \boldsymbol{d}_g - \boldsymbol{X}_{(-t)} \beta_{S(-t)} - \boldsymbol{E}_g^{-1} \boldsymbol{f}_g,$$

and the probability $\boldsymbol{\pi}_{St}$ is

$$\pi_{S_t} = \frac{y_t}{y_t + (1-y_t)\sqrt{h_t^{-1}\tau}\exp\left\{-\frac{1}{2}\tau_t m_t^2 + \frac{1}{2} h_t^{-1} z_t^2\right\}},$$

where $y_t = \rho_t r_\pi$ with $r_\pi = a_\pi/(a_\pi + b_\pi)$, and $\boldsymbol{X}_{(-t)}$ and $\boldsymbol{\beta}_{S(-t)}$ denote the $\boldsymbol{X}$ and $\boldsymbol{\beta}_S$ with the element t removed, respectively.

(4) Since a closed-form full conditional for λ_S is not available, update λ_S using a univariate Gaussian random walk.

(5) Update ξ_β using the Auxiliary Gibbs algorithm (Neal 2000).

(6) Update α_β from its conditional distribution.

$$\alpha|\eta,\tilde{k}\begin{cases}\text{Gamma}\,(a_\alpha+k,\beta_\alpha^*) & \text{with probability } p_\alpha\\ \text{Gamma}\,(a_\alpha+k-1,\beta_\alpha^*) & \text{with probability } 1-p_\alpha\end{cases}$$

where

$$b_\alpha^*=\left(\frac{1}{b_\alpha}-\log(\eta)\right)^{-1},\ \text{and}$$

$$p_\alpha=\frac{a_\alpha+k-1}{a_\alpha+k-1+n/b_\alpha^*}$$

Also,

$$n|\alpha,\tilde{k}\text{Beta}\,(\alpha+1,n)\,.$$

(7) Update ξ_λ using the Auxiliary Gibbs algorithm.

(8) Update α_λ using the same procedure in (6) above.

3.2 Inference from MCMC Results

Due to our formulation for the centering distribution of the DP prior on the regression coefficients, our model can estimate the probability of sharp null hypotheses, such as $H_{0,g}$: $\beta_{1,g} = \cdots = \beta_{T^*,g} = 0$ for $g = 1, \ldots, G$. Other hypotheses may be specified, depending on the experimental goals. We estimate these probabilities by simply finding the relative frequency that the hypotheses hold among the states of the Markov chains.

Our prior model formulation also permits inference on clustering of the G objects. Several methods are available in the literature on DP models to estimate the cluster memberships based on posterior samples. (See, for example, Medvedovic and Sivaganesan 2002; Dahl 2006; Lau and Green 2007.) In the examples below we adopt the least-squares clustering estimation of Dahl (2006) which finds the clustering configuration among those sampled by the Markov chain that minimizes a posterior expected loss proposed by Binder (1978) with equal costs of clustering mistakes.

3.3 Hyperparameters Setting

Our recommendation for setting the hyperparameters is based on computing for each object the least-squares estimates of the regression coefficients, the y-intercept, and the mean-squared error. We then set m_μ to be the mean of the estimated y intercepts and p_μ to be the inverse of their variances. We also use the method of moments to set (a_τ, b_τ). This requires solving the following two equations:

$a_\tau b_\tau$=mean of variance of least−squares regression coefficients

$a_\tau b_\tau^2$=sample variance of variances of least−squares regression coefficients

Likewise, a_λ and b_λ are set using the method of moments estimation, assuming that the inverse of the mean-squared errors are random draws from a gamma distribution having mean $a_\lambda b_\lambda$. As for (a_π, b_π) and (a_π, b_π), a specification such that $(r_\pi E[\rho_t])^T = \prod_{t=1}^{T} p(\beta_{t,g}=0) = 0.50$ is recommended if there is no prior information available.

We refer to these recommended hyperparameter settings as the method of moments (MOM) settings. The MOM recommendations are based on a thorough sensitivity analysis we performed on all the hyperparameters using simulated data. Some results of this simulation study are described in Section 4.1.

4 Applications

We first demonstrate the performance in a simulation study and then apply our method to gene expression data analysis.

4.1 Simulation Study

Data Generation—In an effort to imitate the structure of the microarray data experiment examined in the next section, we generated 30 independent datasets with 500 objects measured at two treatments and three time points, having three replicates at each of the six treatment combinations. Since the model includes an object-specific mean, we set $\beta_{6,g} = 0$ so that the treatment index t ranges from 1 to 5.

We simulated data in which the regression coefficients $\boldsymbol{\beta}$ for each cluster is distributed as described in Table 1. Similarly, the three pre-defined precisions $\lambda_1 = 1.5$, $\lambda_2 = 0.2$ and $\lambda_3 = 3.0$ are randomly assigned to each of the 180, 180, and 140 objects (total 500 objects).

Sample-specific means μ_g were generated from a univariate normal distribution with mean 10 and precision 0.2. Finally, each vector $\boldsymbol{d}_g$ was sampled from a multivariate normal distribution with mean $\mu_g \boldsymbol{j} + \boldsymbol{X}\boldsymbol{\beta}_g$ and precision matrix $\lambda_g \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix.

We repeated the procedure above to create 30 independent datasets. Our interest lies in testing the null hypothesis $H_{0,g}: \beta_{1,g} = \ldots = \beta_{6,g} = 0$. All the computational procedures were coded in Matlab.

Results—We applied the proposed method to the 30 simulated datasets. The model involves several hyperparameters: m_μ, p_μ, a_π, b_π, a_ρ, b_ρ, a_τ, b_τ, a_λ, b_λ, $a_{\alpha\beta}$, $b_{\alpha\beta}$, $a_{\alpha\lambda}$ $b_{\alpha\lambda}$, and We set $(a_\pi, b_\pi) = (1, 0.15)$ and $(a_\rho, b_\rho) = (1, 0.005)$. The prior probability of the null hypothesis (i.e., that all the regression coefficients are zero) for an object is about 50%, which is $(r_\pi * E[\rho])^5$ with $r_\pi = a_\pi/(a_\pi + b_\pi)$ and $E[\rho] = a_\rho/(a_\rho + b_\rho)$, product of Bernoulli random variables across the T treatment conditions each having success probability. We calculated the MOM recommendations from Section 3.3 to set (a_τ, b_τ) and (a_λ, b_λ). These recommendations for the hyperparameters are based on the sensitivity analysis described later in the paper. We somewhat arbitrarily set $(a_{\alpha\beta}, b_{\alpha\beta}) = (5, 1)$ and $(a_{\alpha\beta}, b_{\alpha\beta}) = (1, 1)$, so that prior expected numbers of clusters are about 24 and 7 for the regression coefficients and precisions, respectively. We show the robustness of the choice of those parameters in the

later section. For each dataset, we ran two Markov chains for 5,000 iterations and different starting clustering configurations.

A trace plot of the number of clusters of β from the two different starting stages for one of the simulated datasets, as well as a similar plot for λ, is shown in Figure 1. Similar trace plots of generated $_{\alpha\beta}$ and $_{\alpha\lambda}$ are shown in Figure 2. They do not indicate any convergence or mixing problems. The other datasets also had plots indicating good mixing. For each chain, we discarded the first 3,000 iterations for a burn-in and pooled the results from the two chains.

Our interest in the study is to see whether there are changes between the two groups within a time point and across time points. Specifically, we considered the null hypothesis that all regression coefficients are equal to zero: for $g = 1, \ldots, 500$,

$$H_{0,g}:\beta_{1,g}=\ldots=\beta_{6,g}=0$$

$$H_{a,g}:\beta_{t,g}: \neq 0 \text{ for some } t=1,\ldots,6.$$

We ranked the objects by their posterior probabilities of alternative hypotheses $H_{a,g}$, which equal $1-\Pr(H_{0,g}|\text{data})$. A plot of the ranked posterior probability for each object is shown in Figure 3.

Bayesian False Discovery Rate—Many multiple testing procedures seek to control some type of a false discovery rate (FDR) at a desired value. The Bayesian FDR (Genovese and Wasserman 2003; Müller et al. 2004; Newton et al. 2004) can be obtained by

$$\widehat{FDR}(c)=\frac{\sum_{g=1}^{G} D_g\left(1-v_g\right)}{\sum_{g=1}^{G} D_g}$$

where $v_g = Pr(H_{a,g}|\text{data})$ and $D_g = I(v_g > c)$. We reject $H_{0,g}$ if the posterior probability v_g is greater than the threshold c. The optimal threshold c can be found to be a maximum value of c in the set of $\left\{c:\widehat{FDR}(c) \le \alpha\right\}$ with pre-specified error rate α. We averaged the Bayesian FDRs from the 30 simulated datasets. The optimal threshold, on average, is found to be 0.7 for an Bayesian FDR of 0.05. The Bayesian FDR has also been compared with the true proportion of false discoveries (labeled as "Realized FDR" in the plot) and is displayed in Figure 4. In this simulation, our Bayesian approach is slightly anti-conservative. As shown in Dudoit et al. (2008), anti-conservative behavior in FDR controlling approaches is often observed for data with high correlation structure and a high proportion of true null hypotheses.

Comparisons with Other Methods—We assessed the performance of the proposed method by comparing with three other methods, a standard Analysis of Variance (ANOVA), the SIMTAC method of Dahl et al. (2008), and LIMMA (Smyth 2004). The LIMMA procedure is set in the context of a general linear model and provides, for each gene, an *F*-statistic to test for differential expression at one or more time points. These *F*-statistics were used to rank the genes. The SIMTAC method uses a modeling framework similar to the one we adopt but it is not able to provide estimates of probabilities for $H_{0,g}$ since its posterior

density is continuous. We used the univariate score suggested by Dahl et al. (2008) which captures support for the hypothesis of interest, namely $q_g = \sum_{t=1}^{T} \beta_{t,g}^2$. For the ANOVA procedure, we ranked objects by their p-values associated with $H_{0,g}$. Small p-values indicate little support for the $H_{0,g}$.

For each of the 30 datasets and each method, we ranked the objects as described above. These lists were truncated at 1, 2, …, 200 samples. At each truncation, the proportions of false discoveries are computed and averaged over the 30 datasets. Results are displayed in Figure 5. It is clear that our proposed method exhibits a lower proportion of false discoveries and that performances are substantially better than ANOVA and LIMMA and noticeably better than the SIMTAC method.

Sensitivity Analysis—The model involves several hyperparameters: m_μ, p_μ, a_π, b_π, a_ρ, b_ρ, a_τ, b_τ, a_λ, b_λ, $a_{\alpha\beta}$, $b_{\alpha\beta}$, $a_{\alpha\lambda}$, $b_{\alpha\lambda}$, and m_t. In order to investigate the sensitivity to the choice of these hyperparameters, we randomly selected one of the 30 simulated datasets for a sensitivity analysis.

We considered ten different hyperparameter settings. In the first scenario, called the "MOM" setting, we used all the MOM estimates of the hyperparameters and $(a_\pi, b_\pi) = (1,0.15)$, $(a_\rho, b_\rho) = (1,0.005)$, and $(a_{\alpha\beta}, b_{\alpha\beta}) = (5,1)$. The other nine scenarios with change in one set of parameters given all other parameters set same as in the first scenario were:

- **i.** $(a_\pi, b_\pi) = (15, 15)$, so that $p(\beta_{t,g} = 0) = 0.50$.
- **ii.** $(a_\pi, b_\pi) = (1, 9)$, so that $p(\beta_{t,g} = 0) = 0.10$.
- **iii.** $(a_\rho, b_\rho) = (1, 2)$, so that $E[r_\pi \rho_t] = 0.25$.
- **iv.** $(a_\tau, b_\tau) = (1, 0.26)$, to have smaller variance than MOM estimate.
- **v.** $(a_\tau, b_\tau) = (1, 0.7)$, to have larger variance than MOM estimate.
- **vi.** $(a_\lambda, b_\lambda) = (1, 0.5)$, to have smaller variance than MOM estimate.
- **vii.** $(a_\lambda, b_\lambda) = (1, 3)$, to have larger variance than MOM estimate.
- **viii.** $(a_{\alpha\beta}, b_{\alpha\beta}) = (25, 1)$, to have $E[\alpha_\beta] = 25$, so that prior expected number of clusters is about 77.
- **ix.** $(a_{\alpha\beta}, b_{\alpha\beta}) = (1, 1)$, to have $E[\alpha_\beta] = 1$, so that prior expected number of clusters is about 7.

We set $m_t = 0$. Also, the mean m_μ of the distribution of μ was set to the estimated least-squares intercepts and the precision p_μ to the precision of the estimated intercepts. An identity matrix was used for $\mathbf{M}$ since we assume independent sampling. We fixed $\alpha_\lambda = 1$ throughout the sensitivity analysis. We expect similar sensitivity result of the parameter as one for α_β. We ran two MCMC chains with different starting values; one chain started from one cluster (for both $\boldsymbol{\beta}$ and λ) and the other from G clusters (for both). Each chain was run for 5,000 iterations.

We assessed the sensitivity of the hyperparameter settings in two ways. Figure 6 shows that the proportion of false discoveries is remarkably consistent across the ten different hyperparameter settings. We also identified, for each hyperparameter setting, the 50 objects most likely to be "differentially expressed". In other words, those 50 have the smallest probability for the hypothesis H_0. Table 2 gives the number of common objects among all the pairwise intersections from the various parameter settings. These results indicate a high degree of concordance among the hyperparameter scenarios. We are confident in

recommending, in the absence of prior information, the use of the MOM estimates for (a_τ, b_τ) and (a_λ, b_λ) and to choose (a_π, b_π) and (a_ρ,b_ρ) such that $p(\beta_{t,g} = 0) = 0.50$. The choice for ($a_{\alpha\beta}$, $b_{\alpha\beta}$) does not make a difference in the results.

4.2 Gene expression study

We illustrate the advantage of our method in a microarray data analysis. The dataset was used in Dahl and Newton (2007). Researchers were interested in the transcriptional response to oxidative stress in mouse heart muscle and how that response changes with age. The data has been obtained in two age groups of mice; Young (5-month old) and Old (25-month old) which were treated with an injection of paraquat (50mg/kg). Mice were killed at 1, 3, 5 and 7 hours after the treatment or were killed without having received paraquat (called baseline). So, the mice yield independent measurments, rather than repeated measurements. Gene expressions were measured 3 times at all treatments. Originally, gene expression was measured on 10,043 probe sets. We randomly select $G = 1,000$ genes out of 10,043 to reduce computation time. We also choose the first two treatments, baseline and 1 hour after injection from both groups since it is often of interest to see if gene expressions have been changed within 1 hour after injection. Old mice at baseline were designated as a reference treatment. While the analysis is not invariant to the choice of the reference treatment, we show in Section 5 that the results are robust to the choice of the reference treatment. The data was background-adjusted and normalized using the Robust Multichip Averaging (RMA) method of Irizarry et al. (2003).

Our two main biological goals are to identify genes which either are:

1. Differentially expressed in some way across the four treatment conditions, i.e., genes having small probability of $H_{0,g} : \beta_{1,g} = \beta_{2,g} = \beta_{3,g} = 0$, or
2. Similarly expressed at baseline between old and young mice, but Differentially expressed 1 hour after injection, i.e. genes having large probability of $H_{a,g} : |\beta_{1,g} - \beta_{3,g}| = 0 \;\&|\beta_{2,g} - \beta_{4,g}| > c$, for some threshold c, such as 0.1.

Assuming that information on how many genes are Differentially expressed is not available, we set a prior on π by defining (a_π, b_π) = (10, 3) and (a_ρ, b_ρ) = (100, 0.05) which implies a belief that about 50% of genes are Differentially expressed. We set ($a_{\alpha\beta}$, $b_{\alpha\beta}$) = (5, 5) and ($a_{\alpha\lambda}$, $b_{\alpha\lambda}$) = (1, 1) so that the expected numbers of clusters are 93 and 8 for the regression coefficients and precisions, respectively. Other parameters are estimated as we recommended in the simulation study. We ran two chains starting at two different initial stages: (*i*) all the genes being together and (*ii*) each having its own cluster. The Markov chain Monte Carlo (MCMC) sampler was run for 10,000 iterations with the first 5,000 discarded as burn-in. Figure 7 shows trace plots of the number of clusters for both regression coefficients and precisions. The plots do not indicate convergence or mixing problems. The least-squares clustering method found a clustering for the regression coefficients with 14 clusters and a clustering for the precisions with 11 clusters.

There were six large clusters for β with size more than 50. Those clusters included 897 genes. The average gene expressions for each one of the six clusters are shown in Figure 8(a). The *y*-axis indicates the average gene expressions, and the *x*-axis indicates the treatments. Each cluster shows its unique profile. We found one cluster of 18 genes with all regression coefficients equal to zero (Figure 8(b)).

For hypothesis testing, we ranked genes by calculating posterior probabilities for the genes least supportive of the null hypothesis, $H_{0,g} : \beta_{1,g} = \beta_{2,g} = \beta_{3,g} = \beta_{4,g} = 0$. We listed the fifty genes that were least supportive of the hypothesis $H_{0,g}$. Figure 9 shows the heatmap of those fifty genes.

Finally, in order to identified genes following the second hypothesis of interest $H_{a,g}: |\beta_{1,g} - \beta_{3,g}| = 0 \;\&\; |\beta_{2,g} - \beta_{4,g}| > 0.1$, we similarly identified the top fifty ranked genes. For this hypothesis, our approach clearly finds genes following the desired pattern, as shown in Figure 10.

5 Discussion

We have proposed a semiparametric Bayesian method for random effects models in the context of multiple hypothesis testing. A key feature of the model is the use of a spiked centering distribution for the Dirichlet process prior. Dirichlet process mixture models naturally borrow information across similar observations through model-based clustering, gaining increased power for testing. This centering distribution in the DP allows the model to accommodate the estimation of sharp hypotheses. We have demonstrated via a simulation study that our method yields a lower proportion of false discoveries than other competitive methods. We have also presented an application to microarray data where our method readily infers posterior probabilities of genes being Differentially expressed.

One issue with our model is that the results are not necessarily invariant to the choice of the reference treatment. Consider, for example, the gene expression analysis of Section 4.2 in which we used the group (Old, Baseline) as the reference group. To investigate robustness, we reanalyzed the data using (Young, Baseline) as the reference group. We found that the rankings between two results are very close to each other (Spearman's correlation = 0.9937762, Figure 11).

Finally, as we mentioned in the Section 2.1, our current model can easily accommodate covariates by placing them in the **X** matrix. Such covariates might include, for example, demographic variables regarding the subject or environmental conditions (e.g., temperature in the lab) that affect each array measurement. Adjusting for such covariates has the potential to increase the statistical power of the tests.

Acknowledgments

Marina Vannucci is supported by NIH/NHGRI grant R01HG003319 and by NSF award DMS-0600416. The authors thank the Editor, the Associated Editor and the referee for their comments and constructive suggestions to improve the paper.

1 Appendix

1.1 Full Conditional for Precision

$$\begin{aligned} p(\tau|\mathbf{d},\beta,\lambda) &\propto p(\tau)\prod_{S\in\xi} p(\beta_S|\pi,\tau)\, p(\mathbf{d}_S|\beta_S,\lambda_S,\pi,\tau) \\ &= \left\{\prod_{t=1}^{T} p(\tau_t)\right\}\left\{\prod_{S\in\zeta_t}\prod_{t=1}^{T} p(\beta_{S_t}|\pi_t,\tau_t)\right\} \\ &\propto \prod_{t=1}^{T} p(\tau_t)\left\{\prod_{S\in\zeta_t} N\left(\beta_{S_t}|m_t,\tau_t\right)\right\} \\ &\propto \prod_{t=1}^{T} \tau_t^{a_t+|\zeta_t|/2-1} \exp\left\{-\tau_t\left(\tfrac{1}{b_\tau}+\tfrac{1}{2}\sum_{S\in\zeta_t}\left(\beta_{S_t}-m_t\right)^2\right)\right\} \end{aligned}$$

1.2 Full Conditional for new probability $y_t = \rho_t r_\pi$ of Spike

Note: modified prior $\rho_t r_\pi = p(\beta_t = 0)$ where $r_\pi = a_\pi/(a_\pi + b_\pi)$, thus need a posterior $\rho_t|\text{rest} \propto p(\beta_t = 0|\text{rest})$.

Set $Y_t = r_\pi \rho_t$. Then the distribution of Y_t is

$$\begin{aligned} p(y_t) &= \frac{1}{B(a_\rho, b_\rho)} \left(\frac{y_t}{r_\pi}\right)^{a_\rho - 1} \left(1 - \frac{y_t}{r_\pi}\right)^{b_\rho - 1} \frac{1}{r_\pi} \\ &= \frac{1}{B(a_\rho, b_\rho)} \left(\frac{1}{r_\pi}\right)^{a_\rho + b_\rho - 1} y_t^{a_\rho - 1} (r_\pi - y_t)^{b_\rho - 1}. \end{aligned}$$

Now, we are drawing Y_t, not ρ_t from their conditional distributions: for t,

$$p(y_t|\text{rest}) \propto p(y_t)\ y_t^{\Sigma_{S \in \xi} I(\beta_{St} = 0)} (1 - y_t)^{\Sigma_{S \in \xi} I(\beta_{St} \neq 0)},$$

which is not of known form of distribution. Once we draw samples of Y_t, then we will get ρ_t as Y_t/r_π. We used a grid-based inverse-cdf method. for sampling Y_t.

1.3 Full Conditional for Regression coefficients

$$\begin{aligned} & p\left(\beta_{St}|\lambda_S, \boldsymbol{d}_S, y_t, \beta_{S(-t)}\right) \propto p(\beta_{St}|y_t) \prod_{g \in S} p\left(\boldsymbol{d}_g|\beta_{St}, \beta_{S(-1)}, \lambda_S\right) \\ = \ & y_t \delta_0(\beta_{St}) \prod_{g \in S} p\left(\boldsymbol{d}_g|\beta_{St}, \beta_{S(-t)}, \lambda_S\right) + (1 - y_t) N(\beta_{St}|m_t, \tau_t) \prod_{g \in S} p\left(\boldsymbol{d}_g, \beta_{St}\beta_{S(-t)}, \lambda_S\right) \end{aligned}$$

The first part is obvious. Look at the second part. Set $x_t = (X_{1t}, \ldots, X_{Kt})^{\mathrm{T}}$, $\boldsymbol{X}(-t) = (\boldsymbol{x}_1, \cdots . \boldsymbol{x}_{(t-1)}, \boldsymbol{x}_{(t+1)}, \cdots, \boldsymbol{x}_T)$, and $\boldsymbol{\beta}_{S(-t)} = (\beta_{S1}, \cdots, \beta_{S(t-1)}, \beta_{S(t+1)}, \cdots, \beta_{ST})^{\mathrm{T}}$

The second part is proportional to:

$$\exp\left\{-\tfrac{1}{2}\tau_t(\beta_{St} - m_t)^2\right\} \times \exp \left[-\tfrac{1}{2}\left\{\sum_{g \in S} \boldsymbol{D}_g^T \boldsymbol{Q}_g \boldsymbol{D}_g\right\}\right]$$
$$\text{where } \boldsymbol{D}_g = \boldsymbol{d}_g - \boldsymbol{x}_t \beta_{St} - \boldsymbol{X}(-t)\beta_{S(-t)} - \boldsymbol{E}_g^{-1} \boldsymbol{f}_g$$

$$\propto \exp\left\{-\frac{1}{2}\left(\tau_t \beta_{St}^2 - 2\tau_t m_t \beta_{St}\right)\right\} \exp\left\{-\frac{1}{2}\sum_{g \in S}\left(\boldsymbol{x}_t \beta_{St} - \boldsymbol{A}_g\right)^T \boldsymbol{Q}_g \left(\boldsymbol{x}_t \beta_{St} - \boldsymbol{A}_g\right)\right\}$$

$$\propto \exp\left[-\frac{1}{2}\left\{\beta_{St}^2\left(\tau_t + \sum_{g \in S} \boldsymbol{x}_t^T \boldsymbol{Q}_g \boldsymbol{x}_t\right) - 2\beta_{St}\left(m_t \tau_t + \sum_{g \in S} \boldsymbol{x}_t^T \boldsymbol{Q}_g \boldsymbol{A}_g\right)\right\}\right].$$

Therefore, for each t,

$$\beta_{S_t}|\cdot = \begin{cases} 0 & \text{with probability } \pi_{st} \\ \mathcal{N}\left(\frac{m_t\tau_t+\Sigma_{g\in S}\boldsymbol{x}_t^T\boldsymbol{Q}_g\boldsymbol{A}_g}{\tau_t+\Sigma_{g\in S}\boldsymbol{x}_t^T\boldsymbol{Q}_g\boldsymbol{x}_t}, \tau_t+\Sigma_{g\in S}\boldsymbol{x}_t^T\boldsymbol{Q}_g\boldsymbol{x}_t\right) & \text{with probability } 1-\pi_{st}. \end{cases}$$

References

Antoniak CE. Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems. The Annals of Statistics. 1974; 2:1152–1174. 710.

Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatrics. 2001; 17:509–519. 708.

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. 1995; 57:289–300. 708. Series B: Methodological

Berry DA, Hochberg Y. Bayesian Perspectives on Multiple Comparisons. Journal of Statistical Planning and Inference. 1999; 82:215–227. 708.

Binder DA. Bayesian Cluster Analysis. Biometrika. 1978; 65:31–38. 715.

Brown P, Vannucci M, Fearn T. Multivariate Bayesian variable selection and prediction. J. R. Statist. Soc. B. 1998; 60:627–41. 711.

Cai, B.; Dunson, D. Technical report. Department of Statistical Science; Duke University: 2007. Variable selection in nonparametric random effects models. 709

Dahl, DB. Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. In: Do, K-A.; Müller, P.; Vannucci, M., editors. Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press; 2006. p. 201-218.715

Dahl DB, Mo Q, Vannucci M. Simultaneous Inference for Multiple Testing and Clustering via a Dirichlet Process Mixture Model. Statistical Modelling: An International Journal. 2008; 8:23–39. 708, 709, 711, 713, 719, 720.

Dahl DB, Newton MA. Multiple Hypothesis Testing by Clustering Treatment Effects. Journal of the American Statistical Association. 2007; 102(478):517–526. 708, 711.

Do K-A, Müller P, Tang F. A Bayesian mixture model for differential gene expression. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2005; 54(3):627–644. 708.

Dudoit S, Gibert HN, van der Laan MJ. Resampling-Based Empirical Bayes Multiple Testing Procedures for Controlling Generalized Tail Probability and Expected Value Error Rates: Focus on the False Discovery Rate and Simulation Study. Biometrical Journal. 2008; 50:716–744. 719. [PubMed: 18932138]

Dudoit S, Shaffer JP, Boldrick JC. Multiple Hypothesis Testing in Microarray Experiments. Statistical Science. 2003; 18(1):71–103. 708.

Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica. 2002; 12(1):111–139. 708.

Dunson DB, Herring AH, Engel SA. Bayesian Selection and Clustering of Polymorphisms in Functionally-Related gene. Journal of the American Statistical Association. 2008 in press. 709.

Escobar MD, West M. Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association. 1995; 90:577–588. 713.

Genovese, C.; Wasserman, L. Bayesian Statistics 7. Oxford University Press; 2003. Bayesian and Frequentist Multiple Testing; p. 145-161.717

George E, McCulloch R. Variable selection via Gibbs sampling. J. Am. Statist. Assoc. 1993; 88:881–9. 711.

Gopalan R, Berry DA. Bayesian Multiple Comparisons Using Dirichlet Process Priors. Journal of the American Statistical Association. 1998; 93:1130–1139. 708.

Hochberg Y. A Sharper Bonferroni Procedure for Multiple Tests of Significance. Biometrika. 1988; 75:800–802. 708.

Hommel G. A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test. Biometrika. 1988; 75:383–386. 708.

Irizarry R, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed T. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics. 2003; 4:249–264. 724. [PubMed: 12925520]

Lau JW, Green PJ. Bayesian model based clustering procedures. Journal of Computational and Graphical Statistics. 2007; 16:526–558. 715.

Lucas, J.; Carvalho, C.; Wang, Q.; Bild, A.; Nevins, JR.; Mike, W. Sparse Statistical Modelling in Gene Expression Genomics. In: Do, K-A.; Müller, P.; Vannucci, M., editors. Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press; 2006. p. 155-174.711

MacLehose RF, Dunson DB, Herring AH, Hoppin JA. Bayesian methods for highly correlated exposure data. Epidemiology. 2007; 18(2):199–207. 708, 709. [PubMed: 17272963]

Medvedovic M, Sivaganesan S. Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles. Bioinformatrics. 2002; 18:1194–1206. 715.

Müller P, Parmigiani G, Robert C, Rousseau J. Optimal Sample Size for Multiple Testing: The case of Gene Expression Microarrays. Journal of the American Statistical Association. 2004; 99:990–1001. 717.

Neal RM. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics. 2000; 9:249–265. 714.

Newton M, Kendziorski C, Richmond C, Blattner F, Tsui K. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. Journal of Computational Biology. 2001; 8:37–52. 708. [PubMed: 11339905]

Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004; 5:155–176. 708, 719. [PubMed: 15054023]

Scott JG, Berger JO. An Exploration of Aspects of Bayesian Multiple Testing. Journal of Statistical Planning and Inference. 2006; 136:2144–2162. 708.

Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology. 2004; 3(No. 1) Article 3. 709, 719.

Storey J. The optimal discovery procedure: A new approach to simultaneous significance testing. Journal of the Royal Statistical Society. 2007; 69:347–368. 708. Series B

Storey J, Dai JY, Leek JT. The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. Biostatistics. 2007; 8:414–432. 708. [PubMed: 16928955]

Storey JD. A Direct Approach to False Discovery Rates. Journal of the Royal Statistical Society. 2002; 64(3):479–498. 708. Series B: Statistical Methodology

Storey JD. The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-value. The Annals of Statistics. 2003; 31(6):2013–2035. 708.

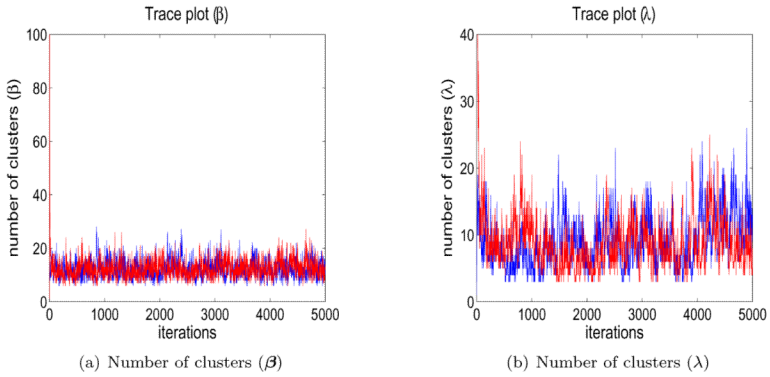
Storey JD, Taylor JE, Siegmund D. Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: a Unified Approach. Journal of the Royal Statistical Society. 2004; 66(1):187–205. 708. Series B: Statistical Methodology

Tibshirani, R.; Wasserman, L. Technical Report 839. Department of Statistics; Carnegie Mellon University: 2006. Correlation-sharing for Detection of Differential Gene Expression. 708

Westfall PH, Wolfinger RD. Multiple Tests with Discrete Distributions. The American Statistician. 1997; 51:3–8. 708.

Westfall, PH.; Young, SS. Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment. John Wiley & Sons; 1993. 708

Yuan M, Kendziorski C. A Unified Approach for Simultaneous Gene Clustering and Differential Expression Identification. Biometrics. 2006; 62:1089–1098. 708. [PubMed: 17156283]



(a) Number of clusters (β) (b) Number of clusters (λ)

Figure 1.
Trace plots of the number of clusters for the regression coefficients and the precisions when fitting a simulated dataset.

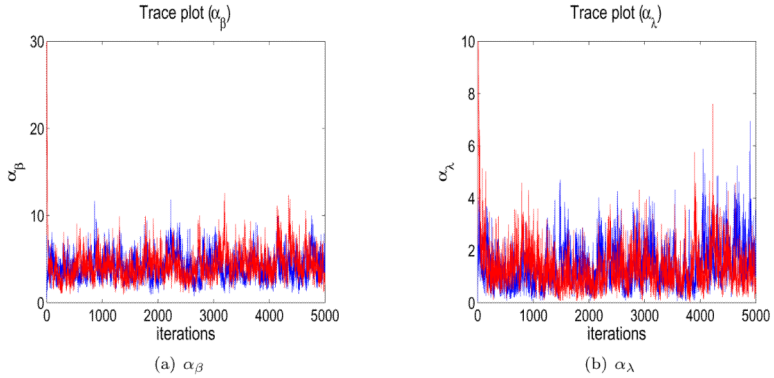


Figure 2.
Trace plots of generated α_β and α_λ when fitting a simulated dataset.

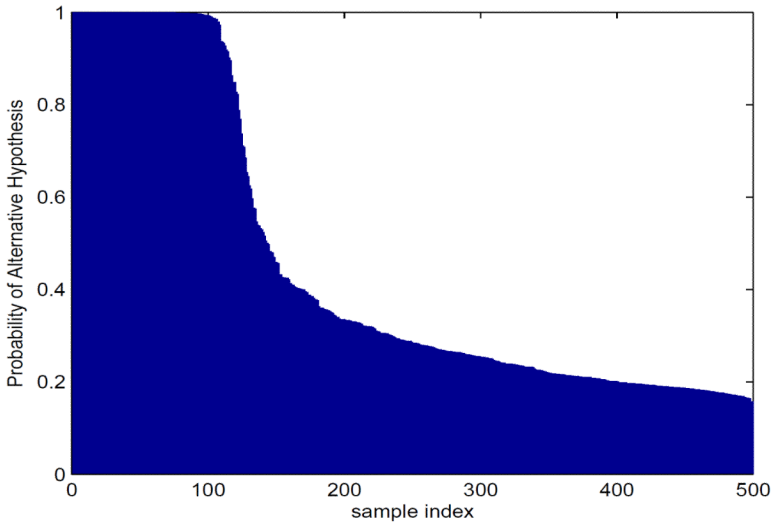


Figure 3.
Probability of the alternative hypothesis (i.e. $1 - \Pr(H_{0,g} : \beta_{1,g} = \ldots = \beta_{6,g} = 0 \mid \text{data})$) for each object of a simulated dataset of 500 objects.

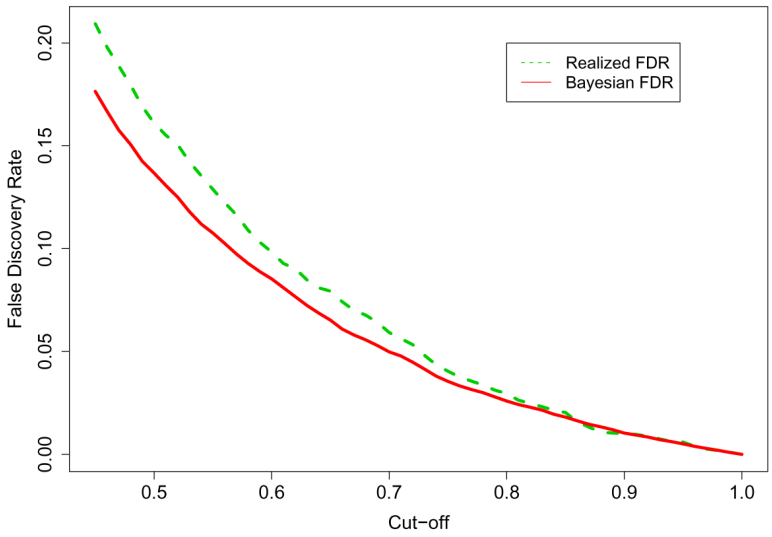


Figure 4.
Plot of proportion of false discoveries and Bayesian FDR averaged over 30 datasets.

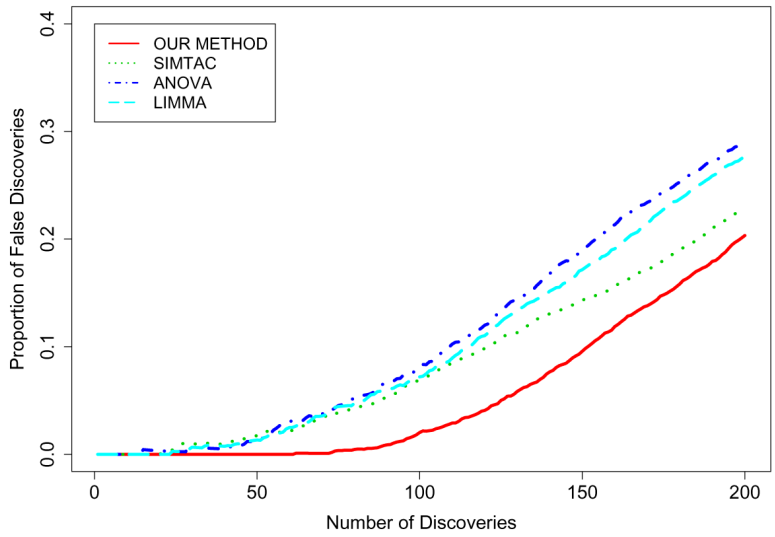


Figure 5.
Average proportion of false discoveries for the three methods based on the 30 simulated datasets

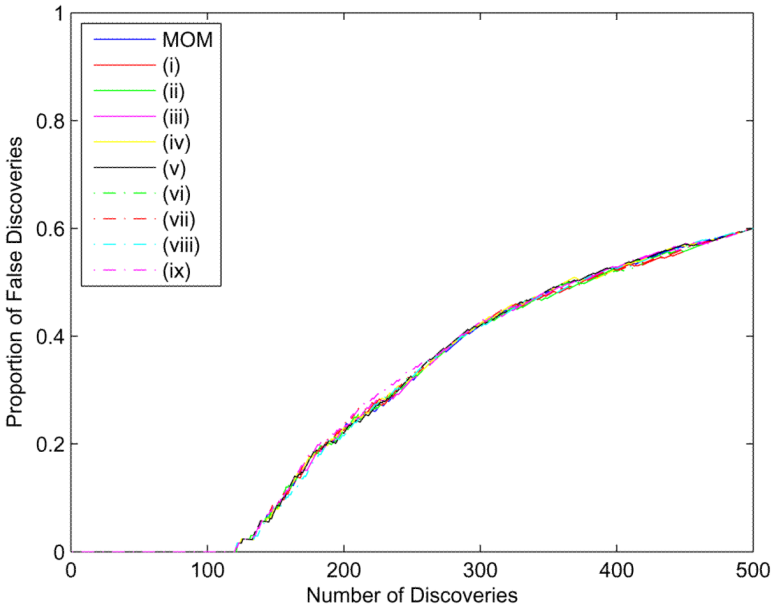
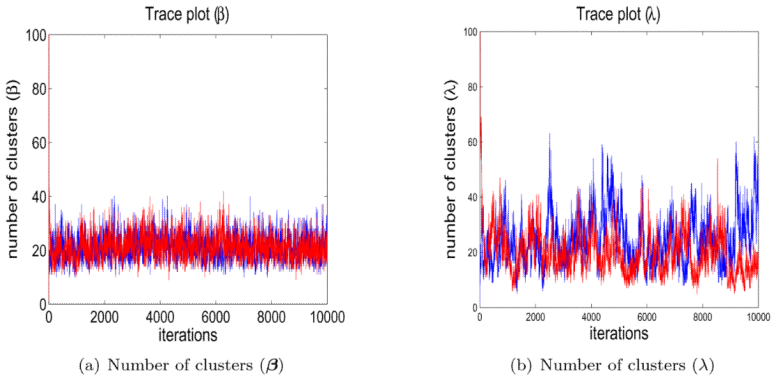
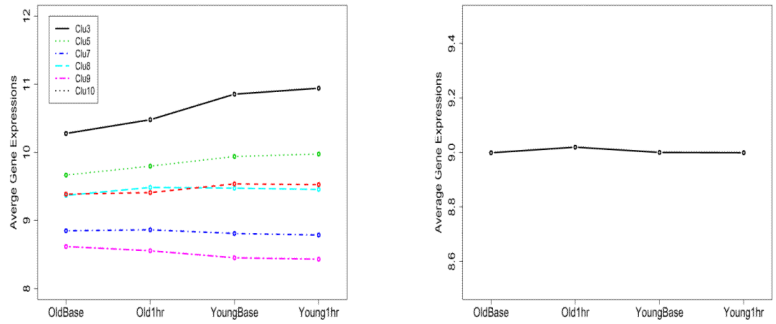


Figure 6.
Proportion of false discoveries under several hyperparameter settings based on one dataset



(a) Number of clusters (β) (b) Number of clusters (λ)

Figure 7.
Trace plots of number of clusters for the regression coefficients and the precisions when fitting the gene expression data.



(a) Large clusters (b) Cluster with estimated $\boldsymbol{\beta} = \mathbf{0}$

Figure 8.
Average expression profiles for (a) six large clusters; (b) cluster with estimated $\boldsymbol{\beta} = \mathbf{0}$.

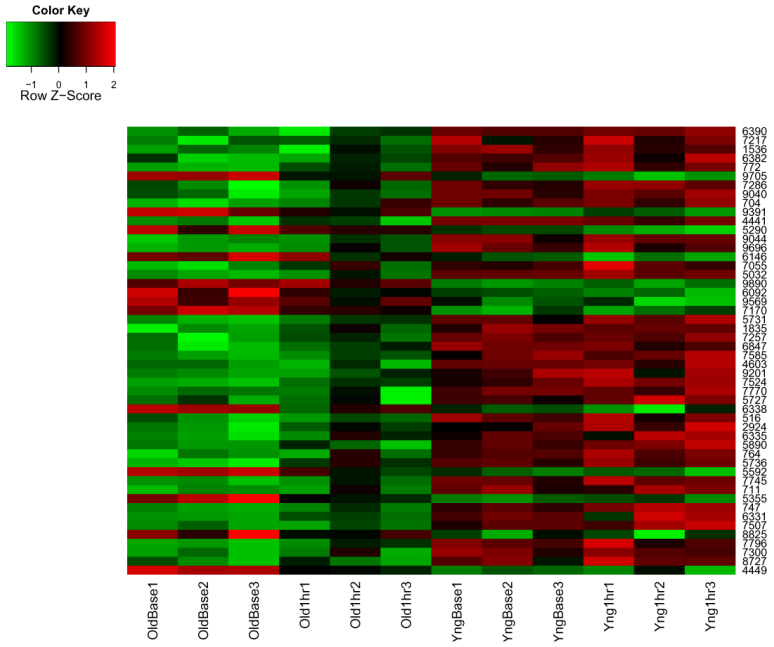
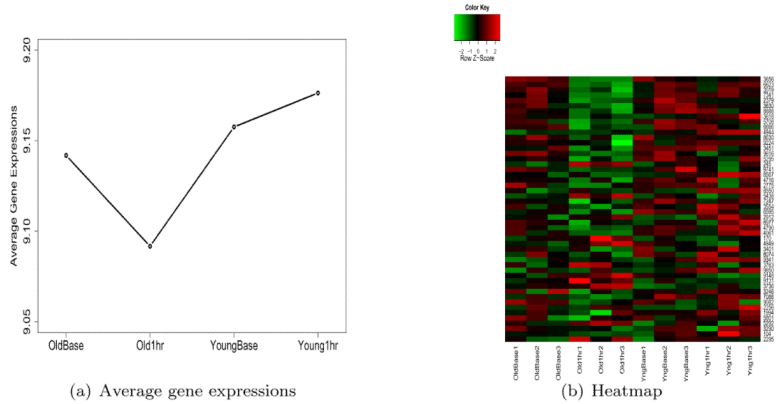


Figure 9.
Heatmap of the 50 top-ranked genes which are least supportive of the assertion that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.



(a) Average gene expressions (b) Heatmap

Figure 10.

(a) Average gene expressions of the 50 top-ranked genes supportive of $|\beta_1 - \beta_3| = 0$ & $|\beta_2 - \beta_4| > 0.1$; (b) Heatmap of those genes

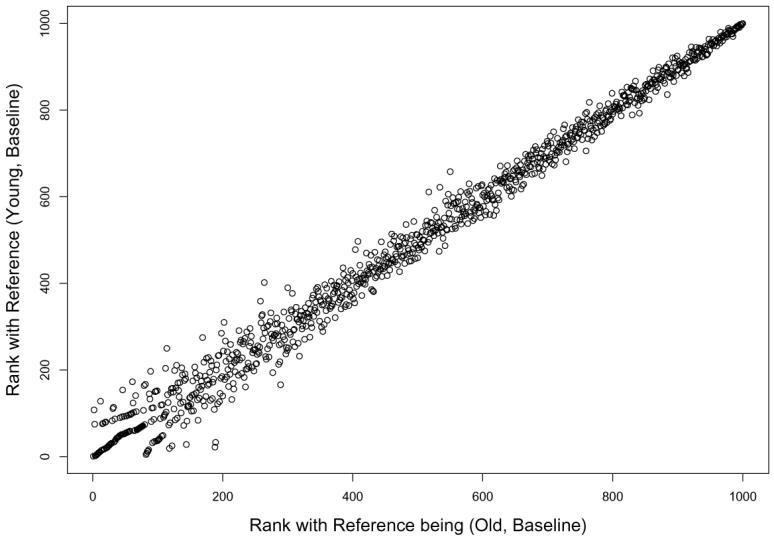


Figure 11.
Scatter plot of rankings of genes resulting from using two reference treatments

Table 1

Schematic for the simulation of the regression coefficients vectors in the first alternative scenario.

Cluster	Size	β_1	β_2	β_3	β_4	β_5
1	300	0	0	0	0	0
2	50	0	0	$\sim N\left(0, \frac{1}{4}\right)$	0	$\sim N\left(0, \frac{1}{4}\right)$
3	50	0	0	0	0	$\sim N\left(0, \frac{1}{4}\right)$
4	25	$\sim N\left(0, \frac{1}{4}\right)$	$\sim N\left(0, \frac{1}{4}\right)$	0	0	0
5	25	0	0	$\sim N\left(0, \frac{1}{4}\right)$	$\sim N\left(0, \frac{1}{4}\right)$	0
6	25	$\sim N\left(0, \frac{1}{4}\right)$	$\sim N\left(0, \frac{1}{4}\right)$	$\sim N\left(0, \frac{1}{4}\right)$	$\sim N\left(0, \frac{1}{4}\right)$	0
7	25	$\sim N\left(0, \frac{1}{4}\right)$	0	$\sim N\left(0, \frac{1}{4}\right)$	0	$\sim N\left(0, \frac{1}{4}\right)$

Table 2

Among the 50 most likely differentially expresed objects, the number in common among the pairwise intersection of the samples identified under the ten hyperparameter settings.

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
MOM (both)	41	37	41	38	39	41	39	39	42
(i)		42	45	45	45	42	45	43	46
(ii)			43	44	43	42	44	42	43
(iii)				45	45	43	45	46	46
(iv)					44	40	47	42	44
(v)						45	44	44	45
(vi)							42	44	45
(vii)								45	44
(viii)									44