Multiple Hypothesis Testing by Clustering Treatment Effects

David B. DAHL and Michael A. NEWTON

Multiple hypothesis testing and clustering have been the subject of extensive research in high-dimensional inference, yet these problems usually have been treated separately. By defining true clusters in terms of shared parameter values, we could improve the sensitivity of individual tests, because more data bearing on the same parameter values are available. We develop and evaluate a hybrid methodology that uses clustering information to increase testing sensitivity and accommodates uncertainty in the true clustering. To investigate the potential efficacy of the hybrid approach, we first study a stylized example in which each object is evaluated with a standard z score but different objects are connected by shared parameter values. We show that there is increased testing power when the clustering is estimated sufficiently well. We next develop a model-based analysis using a conjugate Dirichlet process mixture model. The method is general, but for specificity we focus attention on microarray gene expression data, to which both clustering and multiple testing methods are actively applied. Clusters provide the means for sharing information among genes, and the hybrid methodology averages over uncertainty in these clusters through Markov chain sampling. Simulations show that the hybrid method performs substantially better than other methods when clustering is heavy or moderate and performs well even under weak clustering. The proposed method is illustrated on microarray data from a study of the effects of aging on gene expression in heart tissue.

KEY WORDS: Bayesian nonparametrics; Conjugate Dirichlet process mixture model; Correlated hypothesis test; DNA microarray; Gene expression; Model-based clustering.

1. INTRODUCTION

Research in high-dimensional statistical inference has been motivated in part by problems in genomics, particularly in the analysis of gene expression measured by microarrays. In this domain, two statistical problems are usually treated separately: multiple hypothesis testing and clustering (e.g., Sebastiani, Gussoni, Kohane, and Ramoni 2003). Testing usually aims to detect shifts in the marginal distribution of data on each gene—shifts that emerge when data are obtained under different treatment conditions. Any statistical dependence among data from different genes constitutes a set of nuisance parameters that need to be accommodated for valid testing but are not of primary interest. On the other hand, a main goal of clustering is to group together genes that present highly correlated data; this correlation may reflect underlying biological factors of interest, such as regulation by a common transcription factor. Resolution of these two rather distinct inference problems also must accommodate the high dimensionality of the parameter space and the limited amount of data obtained in each dimension. We propose and study a hybrid method that aims to improve multiple hypothesis testing by explicitly accounting for clusters that share parameter values.

Both multiple hypothesis testing and clustering have been the subject of extensive research in high-dimensional statistical inference. The typical multiple testing scenario involves a method of controlling some type I error rate based on gene-specific test statistics (e.g., Dudoit, Shaffer, and Boldrick 2003). Randomization accommodates among-gene dependence, although the aim of randomization is not to improve the test statistic, but rather to properly control an error rate. The false discovery rate (FDR) is robust to weak among-gene dependencies (van der Laan, Dudoit, and Pollard 2004; Storey, Taylor, and Siegmund 2004), and useful methods are available that from a list of p values produce an error rate–controlled short list of rejected null hypotheses (Storey 2003; Benjamini and Yekutieli 2001; Benjamini and Hochberg 1995). Many methods have been proposed to test for differential gene expression; two that we use for comparison are EBarrays (Kendziorski, Newton, Lan, and Gould 2003) and LIMMA (Smyth 2004). Both of these methods rest on specific hierarchical models of microarray data; neither explicitly accommodates among-gene dependencies. With regard to clustering, model-based methods (Fraley and Raftery 2002; Yeung, Fraley, Murua, Raftery, and Ruzzo 2001; Medvedovic and Sivaganesan 2002) offer certain advantages, although they have had less impact in genomics than more heuristic schemes (Eisen, Spellman, Brown, and Botstein 1998).

A bit of notation takes us to the essence of our proposal. Consider data d_g for each gene g in a genome of G genes. A vector-valued parameter θ_g, together perhaps with some nuisance parameters, indexes a probability distribution for d_g. For example, with expression data in T treatment conditions, θ_g might hold the expected values of expression across the different conditions. The question of whether or not gene g is differentially expressed among conditions can be phrased as the test of a null hypothesis H_{0g} concerning possible values of θ_g. The null hypothesis H_{0g} might enforce equality of the components in the case of two treatments, for example. Let us define true clusters by the rule that two genes, g and g', are in the same cluster if and only if $\theta_g = \theta_{g'}$. A test of H_{0g} using data $d_{g'}$ and d_g has increased power over a test that uses d_g alone, because more data are in play on a common null, but this benefit presumes knowledge that g' should to be clustered with g. A delicate issue is how to achieve power gains when the clusters are estimated.

Section 2 considers a simplified setting, comparing the power of the standard z test to that of a test using some cluster information. Calculations show that the modified procedure is more powerful when the true cluster is estimated sufficiently well.

David B. Dahl is Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *dahl@stat.tamu.edu*). Michael A. Newton is Professor, Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53706 (E-mail: *newton@stat.wisc.edu*). The authors thank Michael Edwards and Tomas Prolla for their generosity with their paraquat data. Marina Vannucci, Gordon B. Dahl, Stephen L. Portnoy, an anonymous associate editor, and anonymous reviewers all provided very helpful comments that substantially improved the manuscript. The first and second authors were supported in part by grants from the National Eye Institute (EY07119) and the National Cancer Institute (R01 CA64364). All computations were performed using the BEMMA software, available at *http://www.stat.tamu.edu/~dahl/software/bemma*.

Journal of the American Statistical Association
June 2007, Vol. 102, No. 478, Theory and Methods
DOI 10.1198/016214507000000211

Extending this to practical methodology for microarray data analysis, we report an approach based on a conjugate Dirichlet process mixture (DPM) model in Section 3. The model contains structures for both marginal mean shifts related to testing as well as dependencies related to clustering. Our proposed model, called BEMMA (Bayesian effects model for microarrays), accommodates uncertainties in true clusters in providing gene-specific hypothesis test results, and the conjugate model formulation leads to reduced complexity of the Markov chain posterior computations (Sec. 3.3). Section 4 presents a simulation study showing that BEMMA can perform substantially better than other methods under various degrees of clustering. Section 5 demonstrates how BEMMA works using microarray data from a study of stress response in mice, and Section 6 concludes with a discussion.

2. DEMONSTRATION OF THE CONCEPT

Power calculations are available in a general, but simplified setting. Consider objects labeled $1, \ldots, n$ with corresponding parameters $\theta_1, \ldots, \theta_n$; null hypotheses $H_{0i}: \theta_i = 0$; and alternative hypotheses $H_{ai}: \theta_i > 0$. For testing, suppose that we have statistics $Z_1, \ldots, Z_n$, which are independent and normally distributed with means $\{\theta_i\}$ and unit variance. Three testing methods use varying degrees of cluster information.

Method 1 applies the standard one-sided level α test to each object i. Thus H_{0i} is rejected when $Z_i > z_\alpha$. Its power at i is $1 - \Phi(z_\alpha - \theta_i)$, where $\Phi(\cdot)$ is the standard normal distribution function and $\Phi(z_\alpha) = 1 - \alpha$.

In contrast, suppose that it is known with certainty which objects cluster with object i and that this knowledge is coded with indicators $c_{ij} = \mathrm{I}\{\theta_i = \theta_j\}$. With

$$S_i = Z_i + \sum_{j \neq i} c_{ij} Z_j, \tag{1}$$

the test statistic $W_i = S_i/\sqrt{n^{(i)}}$ is normally distributed with mean $\sqrt{n^{(i)}}\theta_i$ and unit variance. Here $n^{(i)} = \sum_{j=1}^{n} c_{ij}$ is the size of the cluster containing object i. The level-α test rejects if $W_i > z_\alpha$. Thus the power of method 2 is $1 - \Phi(z_\alpha - \sqrt{n^{(i)}}\theta_i)$, which exceeds that of method 1 unless $n^{(i)} = 1$, in which case both methods are equivalent. Depending on the magnitude of $n^{(i)}$, method 2 can be substantially more powerful than method 1 (especially when θ_i is small).

Methods 1 and 2 represent extreme states of information about clustering. In practice, there is uncertainty about c_{ij} that makes method 2 impossible to implement. But, if clustering uncertainty is low, then some benefits of an approximation, say method 3, over the basic method 1 would be expected. Suppose that we have estimates $\hat{c}_{ij}$ such that

$$\hat{c}_{ij} = \begin{cases} c_{ij} & \text{with probability } 1 - \gamma \\ 1 - c_{ij} & \text{with probability } \gamma, \end{cases}$$

with the clarification that $\hat{c}_{ii} = c_{ii} = 1$. The level of uncertainty is controlled by the error rate γ. Method 3 parallels method 2, but with c_{ij} replaced by $\hat{c}_{ij}$; that is, (1) becomes $\hat{S}_i = Z_i + \sum_{j \neq i} \hat{c}_{ij} Z_j$, which has mean and variance depending on the error rate γ, the true clustering, and $\theta_1, \ldots, \theta_n$. It seems difficult to develop a general and workable test without imposing some additional structure on the problem. We do so explicitly in Section 3, but for a demonstration of concept, here we provide a very stylized model that allows direct power comparisons.

Focus on testing one null $H_{0i}: \theta_i = 0$. Suppose that when $c_{ij} = \mathrm{I}\{\theta_i = \theta_j\} = 0$, θ_j is a random effect drawn independently from a normal distribution with mean 0 and variance τ^2. Also suppose that the induced clustering error $\hat{c}_{ij}$ is independent of statistics $Z_1, \ldots, Z_n$. By the rules of iterated expectation,

$$\mathrm{E}(\hat{S}_i|\theta_i) = \theta_i\big[n^{(i)}(1-\gamma) + \gamma\big]$$

and

$$\begin{aligned}\mathrm{var}(\hat{S}_i|\theta_i) = {} & n^{(i)} - \gamma\big(n^{(i)} - 1\big) + \gamma\big(n - n^{(i)}\big) \\ & + \gamma(1-\gamma)\big[\big(n^{(i)} - 1\big)\theta_i^2 + \big(n - n^{(i)}\big)\tau^2\big] \\ & + \gamma^2\big(n - n^{(i)}\big)\tau^2.\end{aligned}$$

The test statistic $\hat{W}_i = \hat{S}_i/\sqrt{\mathrm{var}(\hat{S}_i|\theta_i)}$ is normally distributed with mean $k\theta_i$ and unit variance, where $k = (n^{(i)}(1-\gamma) + \gamma)/\sqrt{\mathrm{var}(\hat{S}_i|\theta_i)}$, and so the level α test of H_{0i} has power $1 - \Phi(z_\alpha - k\theta_i)$, which exceeds the power of method 1 when $k > 1$. In this stylized example, the clustering error rate γ and nuisance parameters collaborate through the value $k > 0$ to affect power.

Figure 1 illustrates trade-offs in using cluster estimation to improve power. Method 3 is preferred over method 1 only when the cluster error rate is sufficiently low. However, the cluster size $n^{(i)}$ and variance τ^2 affect the comparison; larger clusters can tolerate a higher error rate, and a small τ favors method 3.

We emphasize that these calculations rest on simplifying assumptions to reveal the explicit benefits and weaknesses of hybrid methodology. For instance, we assume that $\hat{c}_{i1}, \ldots, \hat{c}_{in}$ and $Z_1, \ldots, Z_n$ are mutually independent, that a common error rate

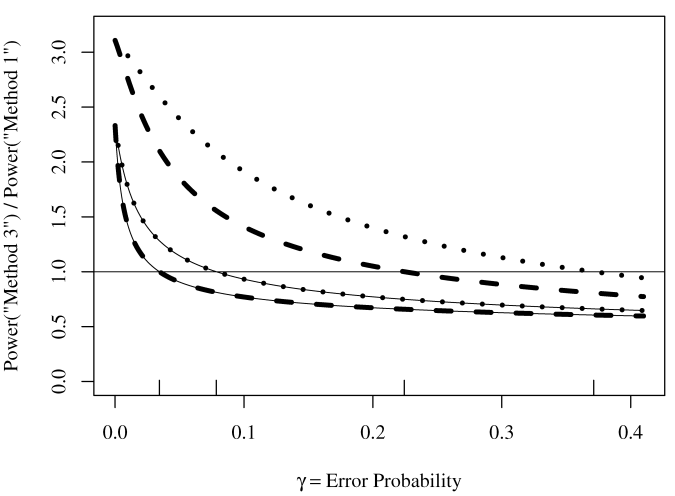

Figure 1. Demonstration of Increased Power When Clustering Is Estimated Well (• •, $n^{(i)} = 30$, $\tau = 1$; ▬, $n^{(i)} = 30$, $\tau = 2$; •—•, $n^{(i)} = 10$, $\tau = 1$; ▬—, $n^{(i)} = 10$, $\tau = 2$). This figure shows the ratio of the power of method 3 (which uses an estimated clustering) and method 1 (z test) versus γ, the error probability that two items are clustered. The four curves correspond to different values for $n^{(i)}$ (the size of the true cluster of object i) and τ [the standard error of the random effects (θ_i)]. For example, when $n^{(i)} = 30$ and $\tau = 2$, method 3 is more powerful than method 1 as long as the error rate is clustering is <.22. The other parameters were set at $\theta_i = .5$ and $n = 500$. Although choosing different values of the parameters affects the curves, the general pattern is preserved.

γ governs both types of errors, and that a simple random-effects formulation accommodates the parameters not under test. Limited as they are, the calculations reveal that testing improvements may be possible if structural assumptions about clustering are incorporated into data analysis. We pursue an instance of this approach in the next section.

3. A MODEL FOR MICROARRAY DATA

3.1 Sampling Distribution

Consider the following sampling distribution:

$$y_{gtr}|\mu_g, \tau_{gt}, \lambda_g \sim \mathrm{N}(y_{gtr}|\mu_g + \tau_{gt}, \lambda_g), \tag{2}$$

where y_{gtr} is a suitably transformed expression measurement in replicate r ($r = 1, \ldots, R_t$) on gene g ($g = 1, \ldots, G$) in treatment condition t ($t = 1, \ldots, T$) and $\mathrm{N}(z|a, b)$ denotes the univariate normal distribution with mean a and precision b (i.e., variance $1/b$) for the random variable z. The parameter μ_g is a gene-specific mean, the gene-specific treatment effects are $\tau_{g1}, \ldots, \tau_{gT}$, and λ_g is a gene-specific sampling precision. The model entails conditional independence of all measurements given the host of gene-level parameters.

Hypotheses about differential expression of gene g involve treatment effects $\tau_{g1}, \ldots, \tau_{gT}$. Simultaneously, the possibility of clustering is accommodated by allowing different genes to have exactly the same treatment effects and precision parameters. Symbolically, genes g and g' are in the same cluster if and only if $(\tau_{g1}, \ldots, \tau_{gT}, \lambda_g) = (\tau_{g'1}, \ldots, \tau_{g'T}, \lambda_{g'})$. A test for differential expression using all genes clustered with gene g may be more sensitive than a test using only data from gene g, as suggested in Section 2. The proposed methodology develops marginal inference on the treatment effects by integrating uncertainty in the cluster structure.

Another set of nuisance parameters are the gene-specific means $\mu_1, \ldots, \mu_G$. These are not related to differential expression across treatments [i.e., are not indexed by conditions in (2)], and we do not use them in defining clusters. These nuisance parameters can be integrated away with respect to some prior, but the resulting model is not conjugate and is computationally prohibitive. We adopt a more pragmatic data-reduction approach. Select a reference treatment, say the first one for convenience. Let d_g be a vector with elements $y_{gtr} - \bar{y}_{g1}$ for $t \geq 2$, where $\bar{y}_{g1}$ is the mean of the reference treatment. Furthermore, let $\tau_g = (\tau_{g2}, \ldots, \tau_{gT})$ be a treatment effect vector and $N = \sum_{t=2}^{T} R_t$ be the dimension of d_g. Then the distribution of d_g does not involve the nuisance parameters $\mu_1, \ldots, \mu_G$,

$$d_g|\tau_g, \lambda_g \sim \mathrm{N}_N(d_g|\mathrm{X}\tau_g, \lambda_g \mathrm{M}), \tag{3}$$

where $\mathrm{N}_c(z|a, b)$ is an c-dimensional multivariate normal distribution with mean vector a and covariance matrix b^{-1} for the random vector z. In addition, M is an $N \times N$ matrix equal to $(\mathrm{I} + \frac{1}{R_1}\mathrm{J})^{-1}$, where I is the identify matrix and J is a matrix of 1's. Finally, X is an $N \times (T-1)$ design matrix whose rows contain all 0's except where the number 1 is needed to pick off the appropriate element of τ_g.

3.2 Clustering Through the Prior

We have defined clustering in terms of ties among $(\tau_1, \lambda_1), \ldots, (\tau_G, \lambda_G)$. This is achieved by assuming that $(\tau_1, \lambda_1), \ldots, (\tau_G, \lambda_G)$ are iid according to an almost-sure discrete, random distribution $F(\tau, \lambda)$, which has a Dirichlet process prior $DP(\eta_0 F_0(\tau, \lambda))$. Such a model is known as a DPM model (see Müller and Quintana 2004 for a review). The mass parameter η_0 controls the degree of clustering; values close to 0 induce many ties, and larger values induce fewer ties. We assume a conjugate centering distribution $F_0(\tau, \lambda)$ to the likelihood in (3),

$$\tau|\lambda \sim \mathrm{N}_{T-1}(\tau|0, \lambda\Psi_0), \qquad \lambda \sim Ga(\lambda|\alpha_0, \beta_0), \tag{4}$$

where $Ga(z|a, b)$ is the gamma distribution with mean a/b for the random variable z and α_0, β_0, and Ψ_0 are hyperparameters whose values must be specified. The Appendix provides an empirical Bayes approach for setting the hyperparameters.

Clustering in terms of ties among $(\tau_1, \lambda_1), \ldots, (\tau_G, \lambda_G)$ is more explicit in an alternative parameterization that uses a set partition $\boldsymbol{\pi} = \{S_1, \ldots, S_q\}$ of $S_0 = \{1, \ldots, G\}$ and a vector of model parameters $\boldsymbol{\phi} = (\phi_{S_1}, \ldots, \phi_{S_q})$, where ϕ_S is associated with cluster S. (The partition $\boldsymbol{\pi}$ satisfies $\bigcup_{S\in\boldsymbol{\pi}} S = S_0$, $S \cap S^* = \emptyset$ for all $S \neq S^*$ and $S \neq \emptyset$ for all $S \in \boldsymbol{\pi}$.) If genes g and g' are clustered [i.e., $(\tau_g, \lambda_g) = (\tau_{g'}, \lambda_{g'})$], then the integers g and g' are in the same cluster S and $\phi_S = (\tau_S, \lambda_S) = (\tau_g, \lambda_g) = (\tau_{g'}, \lambda_{g'})$.

3.3 Sampling From the Posterior

Inference is based on the posterior distribution $p((\tau_1, \lambda_1), \ldots, (\tau_G, \lambda_G)|d_1, \ldots, d_G)$. Much research in Bayesian nonparametrics is devoted to computational techniques for the posterior distribution of DPM models. Quintana and Newton (2000) and Neal (2000) have given reviews and comparisons of methods for fitting DPM models.

Using the set partition representation, the posterior distribution can be factored as $p(\boldsymbol{\phi}|\boldsymbol{\pi}, d_1, \ldots, d_G)p(\boldsymbol{\pi}|d_1, \ldots, d_G)$. The distribution $p(\boldsymbol{\phi}|\boldsymbol{\pi}, d_1, \ldots, d_G)$ is the product over clusters,

$$p(\boldsymbol{\phi}|\boldsymbol{\pi}, d_1, \ldots, d_G) = \prod_{S\in\boldsymbol{\pi}} p(\tau_S|\lambda_S, d_S)p(\lambda_S|d_S),$$

with $d_S = \{d_g \in S\}$ and

$$\begin{aligned} \tau_S|\lambda_S, d_S &\sim \mathrm{N}_{T-1}\big(\tau_S|\Psi_{|S|}^{-1}D_1, \lambda_S\Psi_{|S|}\big) \qquad \text{and} \\ \lambda_S|d_S &\sim Ga\big(\lambda_S|\alpha_{|S|}, \beta_1\big), \end{aligned} \tag{5}$$

where $|S|$ is the number of integers in S and

$$\Psi_k = \Psi_0 + k\mathrm{X}'\mathrm{M}\mathrm{X}, \qquad \alpha_k = \alpha_0 + \frac{kN}{2}, \tag{6}$$

$$\beta_1 = \beta_0 + \frac{1}{2}D_2 - \frac{1}{2}D_1'\Psi_k^{-1}D_1, \tag{7}$$

$$D_1 = \sum_{g\in S} \mathrm{X}'\mathrm{M}d_g, \qquad \text{and} \qquad D_2 = \sum_{g\in S} d_g'\mathrm{M}d_g. \tag{8}$$

Integrating out the model parameters $\boldsymbol{\phi} = (\phi_{S_1}, \ldots, \phi_{S_q})$ reduces the problem to one of running a Markov chain over the posterior clustering distribution $p(\boldsymbol{\pi}|d_1, \ldots, d_G)$. This technique was shown by MacEachern (1994) and MacEachern, Clyde, and Liu (1999) to greatly improve the efficiency of

Gibbs sampling and sequential importance sampling. Given the partition $\boldsymbol{\pi}$, the model parameters can be sampled from (5).

Sampling from the posterior clustering distribution $p(\boldsymbol{\pi}|d_1, \ldots, d_g)$ is computationally challenging but feasible for even large microarray datasets. We recommend using the conjugate Gibbs sampler (MacEachern 1994; Neal 1992) in conjunction with a merge-split sampler of Jain and Neal (2004) or Dahl (2003). (For algorithmic details, we refer the reader to those references as well as to Neal 2000.) It suffices to say that in the context of this model, all of these algorithms rely on

$$p(g \in S) = c_n \exp\bigl(\alpha_{|S|} \log \beta_1 - \alpha_{|S|+1} \log \beta_2\bigr), \tag{9}$$

where

$$\beta_2 = \beta_0 + \frac{1}{2}D_2 + \frac{1}{2}d_g' \mathrm{M} d_g - \frac{1}{2}(\mathrm{X}'\mathrm{M}d_g + D_1)'\Psi_{|S|+1}^{-1}(\mathrm{X}'\mathrm{M}d_g + D_1)$$

and

$$c_n = \frac{\Gamma(\alpha_{|S|+1})}{\Gamma(\alpha_{|S|})}\sqrt{\frac{|\Psi_{|S|}||\mathrm{M}|}{|\Psi_{|S|+1}|(2\pi)^N}}$$

(see Dahl 2004 for details on the derivation).

3.4 Inference

The posterior distribution $p((\tau_1, \lambda_1), \ldots, (\tau_G, \lambda_G)|d_1, \ldots, d_G)$ determines all inferences about expression changes and clustering of the genes. To illustrate inference on differential expression, consider an experiment with $T = 2$ treatments and a differential expression parameter $q_g = (\tau_{g1} - \tau_{g2})^2$. The Bayes estimate of q_g under squared error loss is the posterior mean, which is computable from the Markov chain output. Other parameters are suitable for different experimental designs. For example, one-way layouts might use $q_g = \sum_{i \neq j}(\tau_{gi} - \tau_{gj})^2$, and time course data with $T/2$ periods and two conditions might suggest

$$q_g = \sum_{i \in \{1,3,\ldots,T-1\}} \bigl(\tau_{gi} - \tau_{g(i+1)}\bigr)^2, \tag{10}$$

where odd values of i correspond to the first treatment and indices i and $i + 1$ indicate the same time period. Section 5 gives another score in a different setting. As implemented, BEMMA does not permit the computation of p values or posterior probabilities of point null hypotheses. Instead, inference for altered expression is derived from the posterior distribution of q_g and involves ranking genes according to their evidence of differential expression.

Several methods for clustering inference in DPM models have been proposed. For each $\boldsymbol{\pi}$ among $\boldsymbol{\pi}_{(1)}, \ldots, \boldsymbol{\pi}_{(B)}$ sampled by the Markov chain, an association matrix $\delta(\boldsymbol{\pi})$ of dimension $G \times G$ can be formed whose (i, j) element is $\delta_{i,j}(\boldsymbol{\pi})$, an indicator of whether $(\tau_i, \lambda_i) = (\tau_j, \lambda_j)$. Elementwise averaging yields a matrix of estimates $\hat{p}_{i,j}$ of the pairwise probabilities that objects are clustered. These pairwise probabilities can be used to, for example, identify genes exhibiting similar fluctuations or to find a point estimate of the clustering. Medvedovic and Sivaganesan (2002) used it as a distance matrix in hierarchal 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 $\delta(\boldsymbol{\pi})$ from the pairwise probability matrix

$$\boldsymbol{\pi}_{\mathrm{LS}} = \underset{\boldsymbol{\pi} \in \{\boldsymbol{\pi}_{(1)}, \ldots, \boldsymbol{\pi}_{(B)}\}}{\arg\min} \sum_{i=1}^{G}\sum_{j=1}^{G}(\delta_{i,j}(\boldsymbol{\pi}) - \hat{p}_{i,j})^2. \tag{11}$$

This method minimizes of a posterior expected loss of Binder (1978) with equal costs of clustering mistakes. Section 5 gives examples of clustering inference.

4. SIMULATION STUDY

This section describes a simulation study comparing BEMMA with two other methods for differential gene expression. We find that when the clustering signal is weak or nonexistent, BEMMA still performs well, despite the added burden of searching for cluster structure. In the presence of moderate to heavy clustering, BEMMA is able to take advantage of this structure and deliver improved inference compared with competing methods.

4.1 Synthetic Data

Section 2 suggested that the size of clusters influences the relative power of methods exploiting clustering versus those not using clustering information. Thus the amount of clustering in the simulated datasets will play a key role in how well BEMMA performs. We generated datasets of 1,200 genes under 4 degrees of clustering: heavy clustering (12 clusters of 100 genes per cluster), moderate clustering (60 clusters of 20 genes per cluster), weak clustering (240 clusters of 5 genes per cluster), and no clustering. For each degree of clustering, we generated 30 independent datasets.

The study design is a time course experiment comparing two treatment conditions over three time points. Interest lies in finding genes that are differentially expressed between the two treatment conditions at one or more time points.

The genes in a given cluster share the same model parameters and thus are either all differentially expressed or all equivalently expressed. Clusters of equivalently expressed genes have equal treatment effects for the two treatments within a time point. Clusters of differentially expressed genes have independently sampled treatment effects at one or more of the time points. In all cases, the precision λ for a cluster is drawn from a gamma distribution with mean 1 and variance $1/10$, the treatments effects $\tau_1, \ldots, \tau_6$ for a cluster are drawn independently from a normal distribution with mean 0 and variance $1/(9\lambda)$, and the gene-specific shift μ is drawn from a normal distribution with mean 7 and variance 1.

Regardless of the degree of clustering, each dataset contains 300 genes that are differentially expressed. A third of the differentially expressed clusters have unequal treatment effects at only one time point, a third have unequal treatment effects at two time points, and the remaining third have unequal treatment effects at all three time points. Finally, the observed data are drawn as specified in (2), with the first time point having five replicates per treatment and the other time points having three replicates.

4.2 Differential Gene Expression Results

We applied BEMMA to the simulated datasets, where the hyperparameters were set as recommended in the Appendix. For each dataset, we ran two Markov chains, one with all genes initially together in one cluster and the other with each gene initially in its own cluster. (Sec. 5 discusses the model's sensitivity to changes in the hyperparameters and describes diagnosing convergence of the Markov chain.) After a burn-in period, we ran a Markov chain for 2,000 iterations consisting of a Gibbs scan and a merge-split update of Dahl (2003). Because the autocorrelation time (Ripley 1987; Kass, Carlin, Gelman, and Neal 1998) of the number of clusters was about 7, we effectively had about 300 independent samples per chain. Each gene was ranked by evidence of differential expression using the score in (10).

For comparison purposes, we applied two other methods for detecting differential gene expression to the simulated data: EBarrays (Kendziorski et al. 2003) and LIMMA (Smyth 2004). The EBarrays procedure computes the probability of differential expression for each gene using a mixture model formulation. These probabilities were used to rank the genes. 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.

We used the proportion of false discoveries to compare the three methods. For each of the 30 independent datasets, the methods provided a rankings of the genes in terms of their perception of evidence for differential expression. These lists were truncated at $1, 2, \ldots, 200$ genes. At each truncation, we computed the proportions of false discoveries and averaged them over the 30 datasets. We also computed the standard errors of means, and formed 95% confidence intervals based on the t-distribution. The results are shown in Figure 2.

Under heavy clustering, BEMMA did substantially better than the other methods. This demonstrates that BEMMA is able to exploit the clustering structure to obtain a more sensitive procedure than methods that do not address clustering. Under moderate clustering, BEMMA also outperformed its peers, albeit to a lesser extent because there is less information to exploit with more moderate clustering. Under weak clustering, BEMMA did as well as its peers. Finally, when no clustering was present in the data, the principle motivation of BEMMA was gone, yet BEMMA did not perform much worse than its peers (especially after 100 discoveries).

Whereas these simulation results show the potential for large gains when exploiting clustering, it should be noted that this is just one simulation study. Different parameters values and alternative study designs will yield different relative gains from using BEMMA. In our experience, the most important factor is the size of the true clusters in the data-generating mechanism. BEMMA works best when there is appreciable clustering. In simulation designs where clustering is less prevalent or more difficult to detect (e.g., large within-cluster variances relative to the between-cluster variance), the relative gains of BEMMA will be reduced.

5. EXAMPLE

We demonstrate the proposed method using data from a replicated, multiple-treatment microarray experiment. This example provides a rich context in which to show the flexibility and feasibility of BEMMA. We discuss computational issues, such as burn-in and sampling length, as well as a check on the fitted model and its sensitivity to the hyperparameters.

5.1 Data

In the study under consideration, the researchers were interested in the transcriptional response to oxidative stress in mouse heart muscle and how that response changes with age. Young (5 months) and old (25 months) mice were treated with an injection of paraquat (50 mg/kg). Mice were sacrificed at 1, 3, 5, and 7 hours after paraquat treatment or were sacrificed having not received paraquat (constituting a baseline); thus $T = 10$ experimental conditions were studied. Details of the experiment have been provided by Edwards et al. (2003). All treatments were replicated three times. Gene expression was measured on $G = 10{,}043$ probe sets using high-density oligonucleotide microarrays manufactured by Affymetrix (MG–U74A arrays). The data was background-corrected and normalized using the robust multichip averaging (RMA) method of Irizarry et al. (2003) as implemented in the affy package of BioConductor (Gentleman et al. 2004). Because RMA produces expression values on a log scale, the treatment effects $\tau_{g1}, \ldots, \tau_{gT}$ also should be interpreted on the log scale.

5.2 Nuisance Parameters

The nature of microarray expression data is such that gene-specific means $\mu_1, \ldots, \mu_G$ are nuisance parameters. It is helpful to remove then from the analysis (Sec. 3.1). Figure 3, for example, plots expression values from two different probe sets. They exhibit similar contrasts over time, but clearly have different means. Our method would consider them to probably belong to the same cluster (i.e., to have the same treatment effects and precision), despite the constant shift between them. In fact, these two probe sets correspond to the same gene. Constant differences may be due to aspects of the biology (e.g., mRNA degradation) or hybridization efficiency, rather than to real differences in transcript abundance.

5.3 Sampling From the Posterior

The BEMMA software (*http://www.stat.tamu.edu/∼dahl/bemma*) was used to sample from the posterior distribution. The hyperparameters α_0, β_0, Ψ_0, and η_0 were set according to the Appendix, resulting in the prior and posterior expected number of clusters being 98 (i.e., mass parameter $\eta_0 = 15$).

One iteration of the Markov chain consisted of a Gibbs scan (accounting for more than 97% of the CPU time) and five sequentially allocated merge-split proposals of Dahl (2003). Eight Markov chains were run from one of two extreme starting clusterings: each gene belonging to its own cluster (i.e., 10,043 clusters) or all genes belonging to a single cluster. Using clustering at these polar extremes improves the chances of detecting nonconvergence. Diagnostic plots (not displayed) appeared to indicate that the eight chains converged quickly. More formally, using all eight chains, we applied the Gelman and Rubin

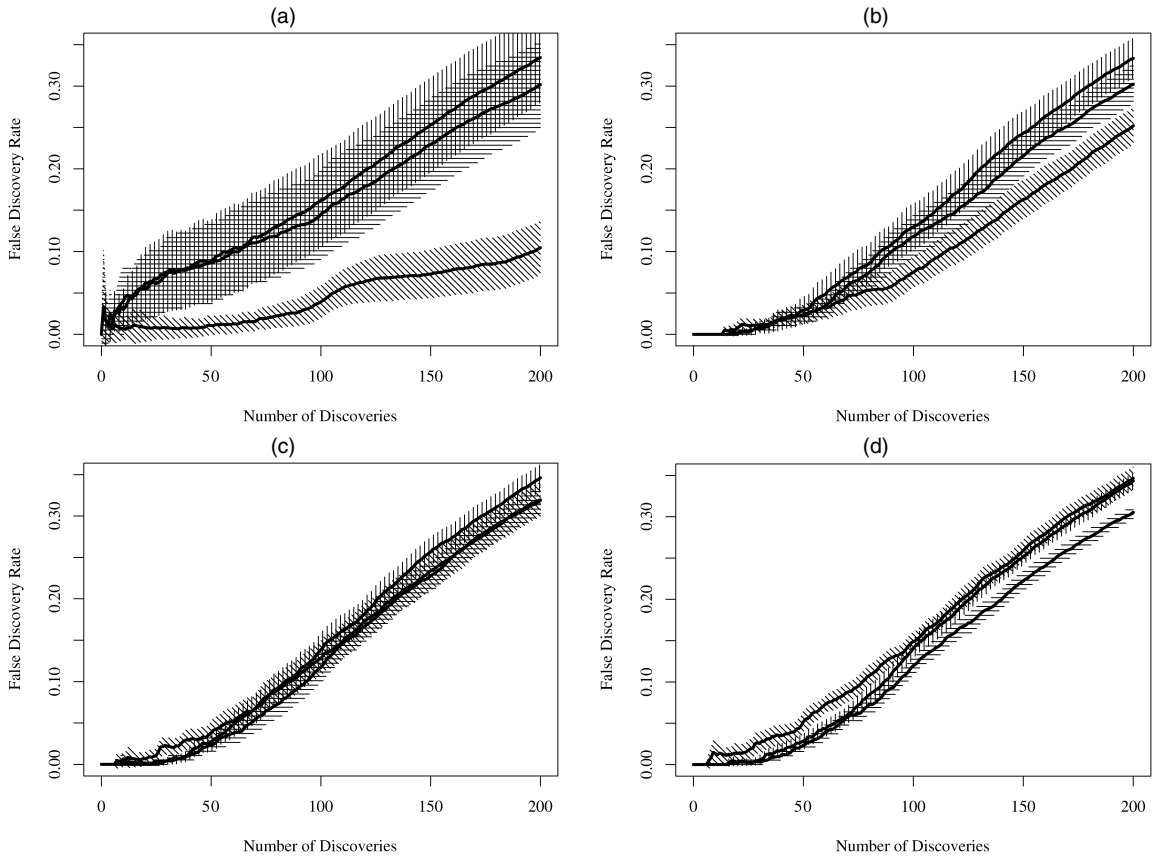

Figure 2. False Discovery Rates for BEMMA and Two Other Methods Under the Four Degrees of Clustering; (a) Heavy, (b) Moderate, (c) Weak, and (d) None (▨ BEMMA; ▤ LIMMA; ▥ EBarrays). The shaded regions give 95% pointwise confidence bands.

(1992) convergence diagnostic for the number of clusters. Figure 4 shows the evolution of this convergence diagnostic with an increasing number of iterations. The diagnostic was close to 1.0 at around 10,000 iterations; thus the first 10,000 iterations from each chain were discarded as burn-in. Applying this burn-in to all 8 chains and then thinning by a factor of 10 yielded a total of 23,994 samples. Figure 5 shows that autocorrelation function of the number of clusters from the thinned samples decayed quickly.

5.4 Differential Expression

Genes with biologically interesting patterns can be identified using the posterior samples. Edwards et al. (2003) observed that immediate early-response genes showed increased expression after paraquat injection. One of the genes that they studied was *zfp36*, the estimated treatment effects and 95% credible intervals of which are shown in Figure 6. Indeed, the expression of *zfp36* significantly increased in response to the paraquat. Further, there was an interaction between age (old vs. young) and

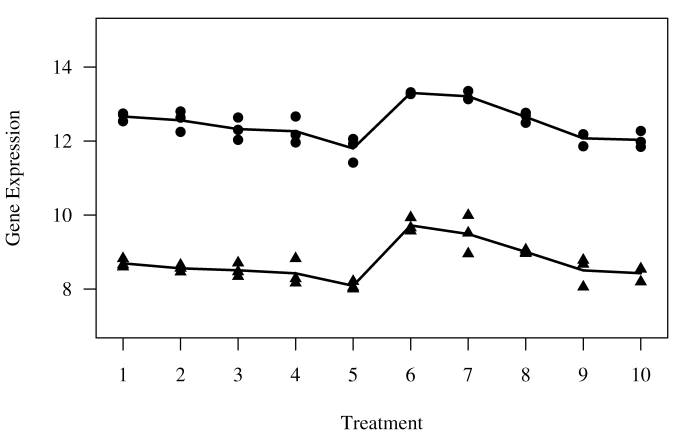

Figure 3. Gene-Specific Shift in Genes Likely to Be Clustered. This figure shows background-corrected and normalized expression of two probe sets exhibiting similar contrasts over time, but clearly having different means. Our method would probably consider them to belong to the same cluster.

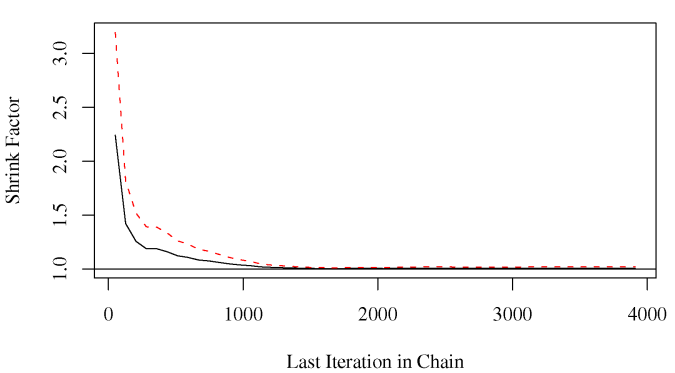

Figure 4. Evolution of the Gelman and Rubin (1992) Convergence Diagnostic for the Number of Clusters (—— median; - - - 97.5%). The diagnostics is close to 1.0 around 10,000 iterations.

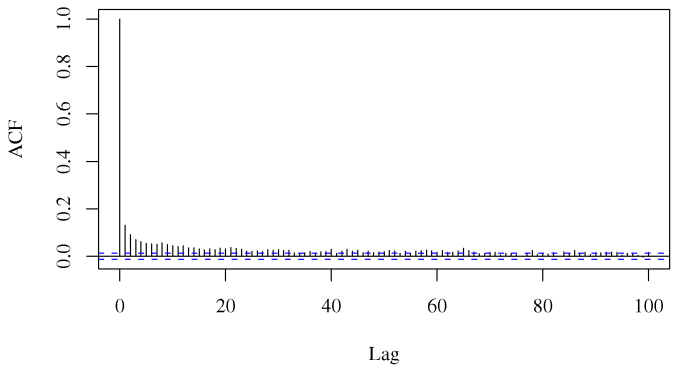

Figure 5. Autocorrelation Function of the Number of Clusters Based on the 1-in-10 Thinned Samples. The dashed lines represent 95% confidence intervals.

time (baseline vs. 1 hour). We found that other immediate early-response genes exhibited similar behavior to that found by Edwards et al. (2003).

Other genes that exhibit a significant age-associated response to the paraquat treatment should have a large posterior mean for the score

$$q = |(\tau_4 - \tau_3) - (\tau_2 - \tau_1)|, \tag{12}$$

where τ_1 is the treatment effect for young mice at baseline, τ_2 is old at baseline, τ_3 is young at 1 hour, and τ_4 is old at 1 hour. Sorting the posterior mean scores, genes showing a large difference between ages at 1 hour compared with the baseline time point can be identified.

5.5 Clustering

The posterior samples also can be used to assess the clustering of genes. The posterior distribution of the number of clusters was nearly normal with mean 98 and standard deviation 4.5. The least squares clustering (Sec. 3.4) had 104 clusters (ranging in size from 1 to 743 genes), with a median cluster size of 57. Among the 500 genes with the largest mean scores in (12), 84% fell into 1 of the 9 clusters from the least squares clustering. The average expression profile for all genes in these nine clusters are displayed in Figure 7.

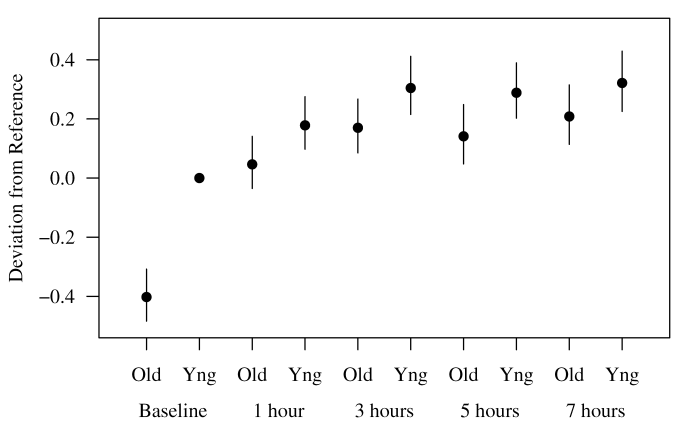

Figure 6. Estimated Treatment Effects for a Probe Set With a Large Interaction Between Age (old vs. young) and Time (baseline vs. 1 hour). The plot also displays 95% credible intervals.

As an example of the analysis of clustering results, we considered cluster I in Figure 7, to which 152 probe sets are associated. The plot shows a persistent, paraquat-independent reduction in old tissue compared with young tissue. Enrichment analysis (e.g., Newton, Quintana, den Boon, Sengupta, and Ahlquist 2007) showed that Gene Ontology categories GO:0042773 and GO:0005746, which are related to mitochondrial electron transport, were enriched for genes in this cluster. This is consistent with related findings in rats (Sandhu and Kaur 2003).

Also consider cluster H, whose genes showed increased expression as a response to the paraquat stress. Note the similarity of Figure 6 for gene *zfp36* (contained in cluster H) and the panel for this cluster in Figure 7. Also contained in cluster H was GADD45β, which was identified by Takekawa and Saito (1998) as encoding for proteins that bind to MAP3K4, a protein known to mediate activation of both p38 and JNK pathways in response to external stimuli. A more detailed discussion of cluster analysis for this example has been given by Dahl (2006).

5.6 Model Checks

As a check of the model fit, we used a multivariate extension of the chi-squared discrepancy measure of Gelman, Carlin, Stern, and Rubin (1995, p. 172),

$$T = \sum_{g=1}^{G} (d_g - \mathrm{X}\tau_g)' \lambda_g \mathrm{M} (d_g - \mathrm{X}\tau_g).$$

The associated posterior predictive p value for this model diagnostic was .65, which is far from 0 or 1, suggesting that the fitted model is consistent with the data.

Although the Appendix provides a default procedure for setting the hyperparameters, it is important to understand how deviations from these recommendations affect posterior inference. To explore robustness, we performed a sensitivity analysis focusing on the prior sample size n_0 and the prior expected number of clusters. (Both of these parameters are described in the App.) For the purposes of this sensitivity analysis, a 1-in-10 subset of the paraquat dataset was obtained, and data from hours 5 and 7 were dropped. Three values for the prior sample size n_0 were considered (.5, 1, and 3), and three values for the prior expected number of clusters were considered (25, 40, and 65). The default values for this dataset are a prior sample size n_0 of 1 and a prior expected number of clusters of 40. For every combination of the these hyperparameters, a Markov chain was run for 50,000 iterations (of which the first 10,000 were discarded). From each chain, the 50 genes with the highest posterior mean for the score in (12) were recorded.

The degree to which the same 50 genes were identified was used as an indication of robustness to the prior. Table 1 shows the percentage of genes identified by both the default model and models under the alternative settings. In all cases, the lists of the top 50 genes had at least 80% of the genes in common. The expected number of clusters (controlled by the mass parameter η_0) seemed to be more robust than the prior sample size n_0, suggesting that this parameter be given extra attention when setting its value.

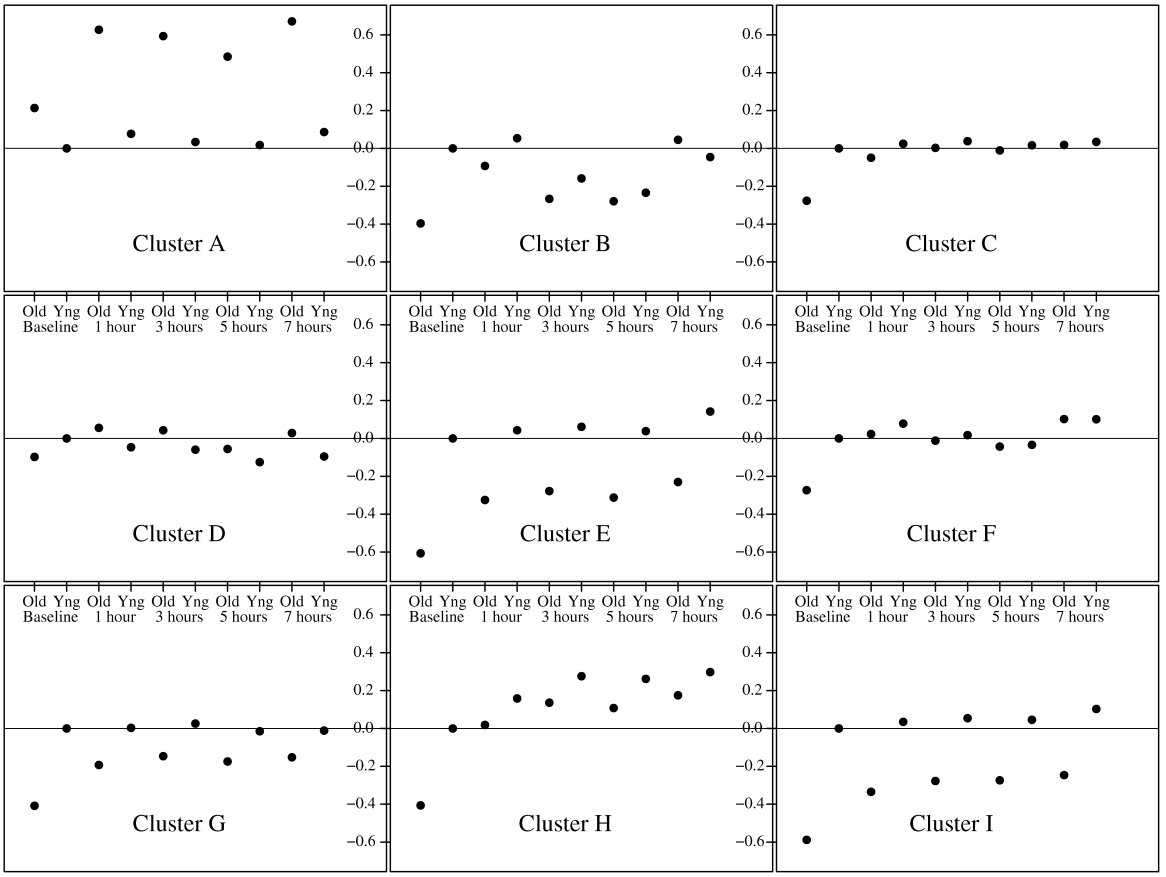

Figure 7. Average Expression Profiles for All Genes in Nine Clusters of Interest. These clusters contain 84% of the 500 genes with the largest mean scores in (12).

6. DISCUSSION

Clustering information can be exploited to improve the sensitivity of correlated hypothesis tests, as we have demonstrated in principle with a toy example (Sec. 2) and empirically in a simulation study of new methodology for microarray expression data. The model behind the BEMMA methodology contains parameters for shifts in marginal expression, as well as parameters for clustering that indicate the genes whose data should to be combined. BEMMA represents an empirical Bayesian methodology. It goes beyond related methods for microarray data by explicitly accommodating among-gene dependence through latent clusters. Genes in the same cluster share values of parameters affecting their expression. We use state-of-the-art Markov chain Monte Carlo schemes to efficiently integrate the unknown clustering. Simulation results show that BEMMA can take advantage of among-gene dependence to improve the identification of differentially expressed genes.

Table 1. Sensitivity of the Ranking of Genes to the Hyperparameters

	Prior expected number of clusters		
Prior sample size	*25*	*40*	*60*
.5	88%	88%	88%
1.0	100%	100%	98%
3.0	78%	78%	80%

NOTE: Under the default settings (a prior sample size of 1.0 and a prior expected number of clusters 40), the 50 genes with the largest posterior mean score in (12) were identified. For other values of the hyperparameters, the top 50 genes were identified. This table shows the percentage of genes that were identified by both the default model and model under the alternative settings.

Our focus has been on taking advantage of clustering to improve testing and gene-specific inferences, but naturally the same machinery can be used to infer the latent cluster structure itself. Using the synthetic data in Section 4, Dahl (2006) compared BEMMA with several clustering methods by way of the adjust Rand index (Rand 1971; Hubert and Arabie 1985). BEMMA was able to estimate the true clustering about as well as MCLUST (Fraley and Raftery 1999, 2002) and much better than hierarchical clustering (Hartigan 1975). Further, the model-based nature of BEMMA allows one to assess the variability in the estimated clustering and provides quantities such as the pairwise probabilities that two genes are clustered.

A common definition in genomics is that clustered genes simply have correlated expression profiles over treatments. But we have defined clustering in terms of equality among latent parameters in a specific nonparametric Bayesian model. We take the point of view that parameters come from an infinite mixture of point distributions and that within a cluster, expression values differ from one another due only to sampling variability. Measurements are (marginally) exchangeable within clusters and independent between clusters. It should be recognized that if two genes are correlated but have different expression parameters, then BEMMA may not put them in the same cluster.

The proposed model is a particular DPM model that is conjugate and thus more computationally efficient than a similar

nonconjugate DPM model. Improved performance in terms of the FDR might be expected if the model were more flexible (by, say, putting prior distributions on the hyperparameters) or did not rely on the specification of a reference treatment. As computing power rapidly expands, such extensions are likely to be practical in the near future.

APPENDIX: SETTING THE HYPERPARAMETERS

Here we give the recommend method for setting the hyperparameters $\eta_0, \alpha_0, \beta_0$, and Ψ_0 of the proposed model. Note that (6) implies that $\Psi_{n+1} = \Psi_n + \mathrm{X}'\mathrm{MX}$ and $\alpha_{n+1} = \alpha_n + \frac{N}{2}$; that is, for each additional observation, Ψ_n and α_n are incremented by $\mathrm{X}'\mathrm{MX}$ and $\frac{N}{2}$. Thus is natural to set the hyperparameter Ψ_0 to $n_0\mathrm{X}'\mathrm{MX}$ and the hyperparameter α_0 to $n_0\frac{N}{2}$, for $n_0 > 0$ representing the number of observations that prior experience is worth. By default, we recommend $n_0 = 1$.

As shown in (2) and (4), the hyperparameters α_0 and β_0 are the shape and rate parameters of the gamma prior distribution for the precision of an observation in a given cluster. We recommend setting α_0 and β_0 such that the mean of this distribution, α_0/β_0, matches a data-driven estimate of the expected precision for a cluster. Equivalently, in terms of the standard deviation, choose α_0 and β_0 so that $\sqrt{\beta_0/\alpha_0}$ matches the estimated standard deviation for a cluster. The software implementation of BEMMA uses the median standard deviation across all probe sets if no value is specified by the user. Because $\alpha_0 = n_0\frac{N}{2}$ (from the previous paragraph), specifying the expected standard deviation implies a value for β_0.

The final hyperparameter to consider is the mass parameter η_0, which affects the distribution on the number of clusters. The mass parameter in DPM models has been well studied (Liu 1996).

From Antoniak (1974), the prior expected number of clusters is $K(G) = \sum_{g=1}^{G} \eta_0/(\eta_0 + g - 1)$. In some DPM model applications, the mass parameter is set to 1.0. This seems overly optimistic for microarray experiments because, for example, it implies a prior belief that there are fewer than 12 clusters in a dataset with 50,000 genes. We use an empirical Bayes approach that sets η_0 such that the posterior expected number of clusters equals the prior expected number of clusters. The software implementation of BEMMA provides this option.

[Received March 2005. Revised January 2007.]

REFERENCES

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174.

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

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188.

Binder, D. A. (1978), "Bayesian Cluster Analysis," *Biometrika*, 65, 31–38.

Dahl, D. B. (2003), "An Improved Merge-Split Sampler for Conjugate Dirichlet Process Mixture Models," Technical Report 1086, University of Wisconsin–Madison, Dept. of Statistics.

——— (2004), "Conjugate Dirichlet Process Mixture Models: Efficient Sampling, Gene Expression, and Clustering," unpublished doctoral thesis, University of Wisconsin–Madison, Dept. of Statistics.

——— (2006), "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model," in *Bayesian Inference for Gene Expression and Proteomics*, eds. K.-A. Do, P. Müller, and M. Vannucci, Cambridge, U.K.: Cambridge University Press, pp. 201–218.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103.

Edwards, M., Sarkar, D., Klopp, R., Morrow, J., Weindruch, R., and Prolla, T. (2003), "Age-Related Impairment of the Transcriptional Responses to Oxidative Stress in the Mouse Heart," *Physiological Genomics*, 13, 119–127.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genomic-Wide Expression Patterns," *Proceedings of the National Academy of Sciences USA*, 95, 14863–14868.

Escobar, M. D. (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Fraley, C., and Raftery, A. E. (1999), "MCLUST: Software for Model-Based Cluster Analysis," *Journal of Classification*, 16, 297–306.

——— (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, New York: Chapman & Hall.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–472.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Li, F. L. C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004), "Bioconductor: Open Software Development for Computational Biology and Bioinformatics," *Genome Biology*, 5, R80.

Hartigan, J. A. (1975), *Clustering Algorithms*, New York: Wiley.

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

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003), "Exploration, Normalization, and Summaries of High-Density Oligonucleotide Array Probe-Level Data," *Biostatistics*, 4, 249–264.

Jain, S., and Neal, R. M. (2004), "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model," *Journal of Computational and Graphical Statistics*, 13, 158–182.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998), "Markov Chain Monte Carlo in Practice: A Roundtable Discussion," *The American Statistician*, 52, 93–100.

Kendziorski, C., Newton, M., Lan, H., and Gould, M. (2003), "On Parametric Empirical Bayes Methods for Comparing Multiple Groups Using Replicated Gene Expression Profiles," *Statistics in Medicine*, 22, 3899–3914.

Liu, J. S. (1996), "Nonparametric Hierarchical Bayes via Sequential Imputations," *The Annals of Statistics*, 24, 911–930.

MacEachern, S. N. (1994), "Estimating Normal Means With a Conjugate-Style Dirichlet Process Prior," *Communications in Statistics, Part B—Simulation and Computation*, 23, 727–741.

MacEachern, S. N., Clyde, M., and Liu, J. S. (1999), "Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation," *The Canadian Journal of Statistics*, 27, 251–267.

Medvedovic, M., and Sivaganesan, S. (2002), "Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles," *Bioinformatrics*, 18, 1194–1206.

Müller, P., and Quintana, F. (2004), "Nonparametric Bayesian Data Analysis," *Statistical Science*, 19, 95–110.

Neal, R. M. (1992), "Bayesian Mixture Modeling," in *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, Dordrecht: Kluwer, pp. 197–211.

——— (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Newton, M., Quintana, F., den Boon, J., Sengupta, S., and Ahlquist, P. (2007), "Random-Set Methods Identify Distinct Aspects of the Enrichment Signal in Gene-Set Analysis," *The Annals of Applied Statistics*, in press.

Quintana, F. A., and Newton, M. A. (2000), "Computational Aspects of Nonparametric Bayesian Analysis With Applications to the Modeling of Multiple Binary Sequences," *Journal of Computational and Graphical Statistics*, 9, 711–737.

Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846–850.

Ripley, B. D. (1987), *Stochastic Simulation*, New York: Wiley.

Sandhu, S., and Kaur, G. (2003), "Mitochondrial Electron Transport Chain Complexes in Aging Rat Brain and Lymphocytes," *Biogerontology*, 4, 19–29.

Sebastiani, P., Gussoni, E., Kohane, I. S., and Ramoni, M. F. (2003), "Statistical Challenges in Functional Genomics," *Statistical Science*, 18, 33–70.

Smyth, G. K. (2004), "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," *Statistical Applications in Genetics and Molecular Biology*, 3, no 1, article 3.

Storey, J. D. (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value," *The Annals of Statistics*, 31, 2013–2035.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society*, Ser. B, 66, 187–205.

Takekawa, M., and Saito, H. (1998), "A Family of Stress-Inducible GADD45-Like Proteins Mediate Activation of the Stress-Responsive MTK1/MEKK4 MAPKKK," *Cell*, 95, 521–530.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004), "Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives," *Statistical Applications in Genetics and Molecular Biology*, 3, no 1, article 15.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001), "Model-Based Clustering and Data Transformations for Gene Expression Data," *Bioinformatics*, 17, 977–987.