Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection

Hemant ISHWARAN and J. Sunil RAO

DNA microarrays open up a broad new horizon for investigators interested in studying the genetic determinants of disease. The high throughput nature of these arrays, where differential expression for thousands of genes can be measured simultaneously, creates an enormous wealth of information, but also poses a challenge for data analysis because of the large multiple testing problem involved. The solution has generally been to focus on optimizing false-discovery rates while sacrificing power. The drawback of this approach is that more subtle expression differences will be missed that might give investigators more insight into the genetic environment necessary for a disease process to take hold. We introduce a new method for detecting differentially expressed genes based on a high-dimensional model selection technique, Bayesian ANOVA for microarrays (BAM), which strikes a balance between false rejections and false nonrejections. The basis of the new approach involves a weighted average of generalized ridge regression estimates that provides the benefits of using shrinkage estimation combined with model averaging. A simple graphical tool based on the amount of shrinkage is developed to visualize the trade-off between low false-discovery rates and finding more genes. Simulations are used to illustrate BAM's performance, and the method is applied to a large database of colon cancer gene expression data. Our working hypothesis in the colon cancer analysis is that large differential expressions may not be the only ones contributing to metastasis—in fact, moderate changes in expression of genes may be involved in modifying the genetic environment to a sufficient extent for metastasis to occur. A functional biological analysis of gene effects found by BAM, but not other false-discovery-based approaches, lends support to this hypothesis.

KEY WORDS: Bayesian analysis of variance for microarrays; False discovery rate; False nondiscovery rate; Heteroscedasticity; Ridge regression; Shrinkage; Variance stabilizing transform; Weighted regression

1. INTRODUCTION

DNA microarray technology allows researchers to measure the expression levels of thousands of genes simultaneously over different time points, different experimental conditions, or different tissue samples. It is the relevant abundance of the genetic product that provides surrogate information about the relative abundance of the cell's proteins. The differences in protein abundance are what characterize genetic differences between different samples. In the preparation of a DNA microarray sample, DNA or RNA molecules labeled with fluorescent dye are hybridized with a library of complementary strands fixed on a solid surface. There are two major branches of chip technologies. Oligonucleotide arrays contain gene-specific sequences, called "probes," about 20 bases long for each gene. The resulting fluorescence intensity from the hybridization process gives information about the abundance of the corresponding sample mRNA (a precursor to the cell's proteins). The other type of array involves complementary DNA (cDNA), which can be spotted on nylon filters or glass slides. Complex mRNA probes are reverse-transcribed to cDNA at two sites for each gene and labeled with red control or green test fluorescent dyes. The ratio of red/green intensities represents the amount of RNA hybridizing at each site. (A good set of references for microarrays is Schena et al. 1998; Schena and Davis 1999; Brown and Botstein 1999; Rao and Bond 2001.)

Although many analysis questions may be of interest, the most commonly posed question asks for the detection of differentially expressing genes between experimental states (e.g., between control samples and treatment samples, or between normal tissue samples and diseased tissue samples). Current approaches involve Bayes and empirical Bayes mixture analysis (Efron, Tibshirani, Storey, and Tusher 2001; Ibrahim, Chen, and Gray 2002; Newton, Kendziorski, Richmond, Blattner, and Tsui 2001), and multiple hypothesis testing approaches with corrections designed to control the expected *false discovery rate* (FDR) using the methods of Benjamini and Hochberg (1995) (see Tusher, Tibshirani, and Chu 2001; Storey 2002; Storey and Tibshirani 2001). The FDR is defined as the false-positive rate among all rejected (null) hypotheses; that is, the total number of rejected hypotheses where the null is in fact true, divided by the total number of rejected hypotheses. Benjamini and Hochberg (1995) provided a sequential p-value method to control the expected FDR. (Note: what they called the FDR is what we refer to as the expected FDR, a slight departure in terminology.) Given a set of independent hypothesis tests and corresponding p values, the method provides an estimate k, such that if one rejects those hypotheses corresponding to $P_{(1)}, P_{(2)}, \ldots, P_{(k)}$, the k-ordered observed p values, then the FDR is on average controlled under some prechosen α level. For convenience, we call this the *BH method.*

ANOVA-based models are another way to approach the problem. The first ANOVA methods were developed to account for ancillary sources of variation when making estimates of relative expression for genes (see, e.g., Kerr, Martin, and Churchill 2000; Wolfinger et al. 2001). More recently, an efficient approach casting the differential detection problem as an ANOVA model and testing individual model effects with FDR corrections was developed by Thomas, Olson, Tapscott, and Zhao (2001). With all of these FDR applications, the methods work well by ensuring that an upper bound is met; however, a side effect is often a high *false nondiscovery rate* (FNR). The FNR is defined as the proportion of nonrejected (null) hypotheses that are incorrect. Genovese and Wasserman (2002a) showed

Hemant Ishwaran is Associate Staff, Department of Biostatistics and Epidemiology Wb4, Cleveland Clinic Foundation, Cleveland, OH 44195 (E-mail: *ishwaran@bio.ri.ccf.org*). J. Sunil Rao is Associate Professor, Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106 (E-mail: *sunil@hal.epbi.cwru.edu*). The authors thank the editor, Frank Samaniego, the two referees, and the associate editor for their time and effort spent in helping to improve this manuscript. The authors also thank Sanford Markowitz, Patrick Leahy, and Petra Platzer of Case Western Reserve University and Keith Wilson and Ghassan Ghandour of EOS Biotechnology for invaluable discussions as well as assistance in running the GeneSpring analyses. J. Sunil Rao was supported by National Institutes of Health grant K25-CA89867.

Journal of the American Statistical Association
June 2003, Vol. 98, No. 462, Theory and Methods
DOI 10.1198/016214503000224

that the BH method cannot simultaneously optimize the expected FDR and the expected FNR, implying that the method has low power. (Our simulations in Sec. 6 also confirm this behavior.) An oracle threshold value was described by Genovese and Wasserman (2002b) that improves on BH in the case where p-value distributions are independent two-point mixtures. Storey (2002) also recognized that power for the BH approach could be improved. Typically, however, there is a price to be paid in terms of increased computation and some subjectiveness. Moreover, the largest relative power gains observed by Storey (2002) will be realized only when large proportions of genes are truly differentially expressed, a property that might not hold in some disease problems, because the number of genes differentially expressed compared with the full dimension are expected to be small.

1.1 Larger Models Versus False Discovery

Our approach to detecting gene expression changes uses a Bayesian ANOVA for microarrays (BAM) technique that provides shrinkage estimates of gene effects derived from a form of generalized ridge regression. This method adapts work by Ishwaran and Rao (2000) for high-dimensional model selection in linear regression problems. A key feature of the BAM method is that its output permits different model estimators to be defined, and each can be tailored to suit the various needs of the user. For example, for analysts concerned with false discovery rates, we show how to construct an estimator that goes after the FDR. Also developed is a simple graphical method based on the amount of shrinkage that can be used to visualize the trade-off between a low FDR and finding more genes. This device can be used to select α (significance level) cutoff values for model estimators. Selecting an appropriate α value is critical to the performance of any method used to detect differentially expressing genes. Simply relying on using preset α values, particularly conventional values used in traditional problems, can be a poor strategy, because such values can sometimes be magnitudes off from optimal ones. Our case study example of Section 7 illustrates how small α can be in practice.

Table 1. Results From the Gene Simulation Model ($\alpha = .05$)

	Detected	*TotalMiss*	*FDR*	*FNR*	*Type I*	*Type II*
Zcut	481	675	.162	.063	.009	.597
FDRmix	416	702	.142	.067	.007	.643
Z-test	1106	792	.406	.039	.049	.343
BH	148	862	.034	.087	.001	.857

As a consequence of shrinkage and model averaging, the BAM method strikes a nice balance between identifying large numbers of genes and controlling the number of falsely identified genes. This kind of property can be of great importance in the search for a colon cancer metastatic signature, a topic that we explore more in Section 7 as one illustration of our method. Currently, very little is known about the genetic determinants of colon cancer metastasis, although it is generally agreed that a genetic signature will be complex. In fitting in with this general philosophy, we work under the hypothesis that genes with large differential expressions may not be the only ones contributing to metastasis—that in fact, more moderate changes in expression of some genes might be sufficient to trigger the process. Proving this hypothesis directly is difficult. A more reasonable surrogate hypothesis conjectures that the genes that show more moderate changes in expression provide a suitable milieu for other metastatic events (accompanied by other genes showing much larger expression differences). This general principle may be at play in many diseases other than colon cancer, and so increased power in detecting differentially expressed genes becomes more important, thus motivating a method like BAM.

To fix ideas about BAM, consider Table 1 and Figure 1, which present results from a gene simulation model involving $p = 10,000$ genes (specific details are provided in Sec. 6).

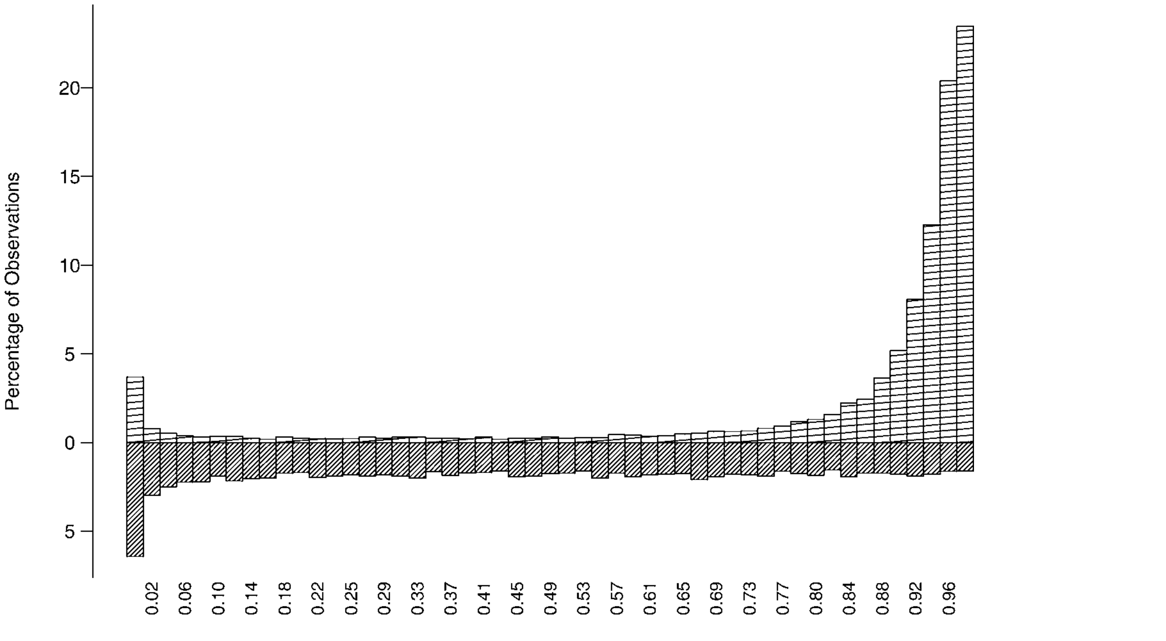

Figure 1. p Values From a Gene Simulation Model Using p = 10,000 Genes With a Mean Group Separation of 1.5 for 10% of Genes and 0 for 90% of Genes. Each gene had five observations per group. The top plot are p values derived from the Zcut estimator, and the bottom plot are p values derived from standard Z-tests.

Figure 1 plots the histogram of the p values from our "Zcut" method (Sec. 3.3) against the p values derived from individual Z-tests that use a pooled estimate for the standard error (Sec. 5.1). Figure 1 shows the effect that shrinkage plays in BAM; here it has the effect of pushing the p-value distribution for Zcut apart, helping to more clearly delineate expressive genes. Differences in procedures can be carefully compared in Table 1, which records the number of genes detected as differentially expressing, the FDR and FNR, and the type I and type II error rates. The *type I error rate* is the proportion of genes rejected given that they are not differentially expressing (the null), whereas the *type II error rate* is the proportion of nonrejected genes given that they are differentially expressing (the alternative). Values are tabulated at an $\alpha = .05$ value. Observe how Zcut leads to a reduced FDR, while at the same time seeks to maintain high power. Table 1 also records the results of the BH method applied to the p values from the Z-tests. Also recorded are the results from the "FDRmix" procedure (Sec. 4), a hybrid BH procedure. Table 1 shows that both BH and FDRmix lead to a smaller number of identified genes than Zcut or Z-test. This is because both procedures are trying to control the FDR which typically results in fewer genes being found significant. Here, the BH method has the lowest FDR, slightly smaller than its target $\alpha = .05$ value. Although FDRmix is far off from the α target value, it does reduce the FDR of Zcut while maintaining power.

Of course, one must be careful when directly comparing FDR and FNR values (or, for that matter, type I and type II error rates) for the different procedures at the same α value, because an α target value means something different for each procedure. Looking at the different rates individually will also not tell us how the procedures perform overall. Thus, to be able to compare procedures on a more equal basis, we have defined an overall measure of performance, "TotalMiss," which is also recorded in Table 1. This is the total number of false rejections and false nonrejections, that is, the number of misclassified genes for a procedure, which can be thought of as a measure of total risk. In terms of the total risk, Table 1 shows that Zcut is the clear winner here. A more detailed study of how TotalMiss varies with α is presented in Section 6. This kind of analysis is important in assessing the robustness of a procedure. Because α values can vary considerably depending on the data, procedures that are robust are those that exhibit uniformly good risk behavior.

1.2 Organization of the Article

The article is organized as follows. Section 2 presents an overview of a stochastic variable selection algorithm that involves shrinkage of ordinary least squares estimators. Section 3 introduces the BAM procedure, along with the Zcut estimator and a shrinkage plot that can be used as a graphical device for setting α values. Theorems 1 and 2 in Section 3.4 discuss the robustness and adaptiveness of the BAM method and provide explanations for the Zcut method (with proofs provided in App. B). Section 4 introduces the FDRmix procedure. Section 5 discusses some more commonly known procedures. Section 6 studies the performance of BAM via simulations and compares it in detail with the more standard methods of Section 5. Section 7 discusses the colon cancer problem and presents a detailed analysis of the data based on BAM. Section 8 concludes with a discussion.

2. PARAMETRIC STOCHASTIC VARIABLE SELECTION

Our approach is to recast the problem of finding differentially expressing genes as a problem of determining which factors are significant in a Bayesian ANOVA model. This is what we call the BAM method. Because one can always rewrite an ANOVA model as a linear regression model, the task of finding expressive genes can be conceptualized as a variable selection problem. This connection allows us to adapt a technique discussed by Ishwaran and Rao (2000) for selecting variables in high-dimensional regression problems called *parametric stochastic variable selection* (P-SVS). Section 3 outlines the details of the BAM approach and how it relates to P-SVS. Here we give a general description of the P-SVS procedure.

The P-SVS method is a hybrid version of the spike and slab models first pioneered by Mitchell and Beauchamp (1988). It is a Bayesian hierarchical approach for model selection in linear regression models,

$$Y_i = x_i^T \beta_0 + \epsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where Y_i is the response value, the ϵ_i are iid normal$(0, \sigma_0^2)$ measurement errors, and $\beta_0 = (\beta_{1,0}, \dots, \beta_{p,0})^T$ is the unknown (true) p-dimensional vector of coefficients for the covariates x_i. To answer the question of which $\beta_{j,0}$ are non-zero, the P-SVS approach works by modeling (1) as a hierarchical model,

$$\begin{aligned}
(Y_i|\beta, \sigma^2, x_i) &\overset{\text{ind}}{\sim} \text{normal}(x_i^T\beta, \sigma^2), \qquad i = 1, \dots, n,\\
(\beta_j|\gamma_j, \tau_j^2) &\overset{\text{ind}}{\sim} \text{normal}(0, \gamma_j\tau_j^2), \qquad j = 1, \dots, p,\\
(\gamma_j|\gamma^*, \lambda) &\overset{\text{iid}}{\sim} (1-\lambda)\delta_{\gamma^*}(\cdot) + \lambda\delta_1(\cdot),\\
(\tau_j^{-2}|a_1, a_2) &\overset{\text{iid}}{\sim} \text{gamma}(a_1, a_2),\\
\lambda &\sim \text{uniform}[0, 1],\\
(\sigma^{-2}|b_1, b_2) &\sim \text{gamma}(b_1, b_2).
\end{aligned} \tag{2}$$

A key feature of the model is the choice of the priors of τ_j^2 and γ_j which are calibrated so that the variance $\nu_j^2 = \gamma_j\tau_j^2$ for a coefficient β_j has a bimodal distribution. A large value for ν_j^2 occurs when $\gamma_j = 1$ and τ_j^2 is large and will induce large values for β_j, thus identifying the covariate as potentially informative. Small ν_j^2 values occur when $\gamma_j = \gamma^*$ (chosen to be a small value). In this case the value for β_j will become near 0, signifying that β_j is unlikely to be informative. The value for λ in (2) controls how likely that γ_j is 1 or γ^*, and thus it controls how many β_j are non-zero and so the complexity of the model. Of course, the choice of hyperparameters for priors are crucial to the behavior of ν_j^2 and hence P-SVS's ability to properly select variables. We use the values for hyperparameters γ^*, a_1, a_2, b_1 and b_2 discussed by Ishwaran and Rao (2000). Note that these values do not need to be tuned for each dataset and can be kept fixed (see Ishwaran and Rao 2000 for details).

Let $\gamma = (\gamma_1, \dots, \gamma_p)^T$. This is a way of encoding models in binary strings (if $\gamma_j = 1$, select β_j). One approach used in spike and slab variable selection looks to the posterior behavior of γ to identify the "best model"; for example, by identifying the γ that occurs with highest posterior probability (George and McCulloch 1993). However, as argued by Ishwaran and Rao (2000), in very large variable selection problems we need to process information differently, because the information contained in γ will be too coarse. (If p is very large, then a high-frequency model may not even be found.) As was argued, variables should be selected by considering the magnitude of their posterior mean values. Motivation to use the posterior mean to select variables stems from its interpretation as an adaptive weighted average of generalized ridge estimates. Such values are reliable, because they are produced by model averaging in combination with shrinkage, two methods known to improve model selection. One can easily see that the posterior mean is a weighted ridge estimate. Let $\beta = (\beta_1, \dots, \beta_p)^T$. From (2), the posterior mean can be written as

$$E(\beta|Y) = \iint \beta\pi(d\beta|\nu^2, \sigma^2, Y)\pi(d\gamma, d\tau^2, d\lambda, d\sigma^2|Y),$$

where $\nu^2 = (\nu_1^2, \dots, \nu_p^2)^T$, $\tau^2 = (\tau_1^2, \dots, \tau_p^2)^T$, and $Y = (Y_1, \dots, Y_n)^T$. Elementary calculations show that

$$(\beta|\nu^2, \sigma^2, Y) \sim \text{normal}(\Sigma^{-1}X^TY, \sigma^2\Sigma^{-1}), \qquad (3)$$

where X is the $n \times p$ design matrix, $\Sigma = \sigma^2\Gamma^{-1} + X^TX$, and Γ is the $p \times p$ diagonal matrix with diagonal values $\nu_j^2 = \gamma_j\tau_j^2$. Notice that the conditional posterior mean for β is

$$E(\beta|\nu^2, \sigma^2, Y) = \Sigma^{-1}X^TY = (\sigma^2\Gamma^{-1} + X^TX)^{-1}X^TY,$$

which is the ridge estimate from a generalized ridge regression of Y on X with weights $\sigma^2\Gamma^{-1}$. Small values for diagonal elements of Γ have the effect of shrinking coefficients. Thus the posterior mean for β can now seen to be a weighted average of ridge shrunken estimates where the adaptive weights are determined from the posteriors of γ, τ^2 and λ.

Remark 1. This shift from high-frequency models selected by γ to models selected on the basis of individual performance of variables was also discussed recently by Barbieri and Berger (2002). These authors showed that the high-frequency model can be suboptimal even in the simplest case when the design matrix is orthogonal. Under orthogonality, the optimal predictive model is not the high-frequency model, but rather the median probability model, defined as the model consisting of variables whose overall posterior probability is greater than or equal to 1/2.

3. BAYESIAN ANOVA FOR MICROARRAYS

The BAM method applies this powerful variable selection device by recasting the microarray data problem in terms of an ANOVA model, and hence as a linear regression model. Here we consider the case in which we have 2 groups of samples; extensions to allow for more groups are discussed in Appendix A. For group l, let $Y_{j,k,l}$ denote the kth expression value, $k = 1, \dots, n_{j,l}$, for gene $j = 1, \dots, p$. Group $l = 1$ corresponds to the control group, and $l = 2$ represents the treatment group. For example, in the colon cancer study of Section 7, the control group represents Duke's B-survivor colon tissue samples, whereas the treatment group are metastasized colon cancer samples. Microarray data, as in our colon cancer study, will often be collected from balanced experimental designs with fixed group sizes $n_{j,1} = N_1$ and $n_{j,2} = N_2$. In such settings, $Y_{1,k,l}, \dots, Y_{p,k,l}$ will typically represent the p gene expressions obtained from a microarray chip for a specific individual k with a tissue type from group l (i.e., either an individual k from the control group or an individual k from the treatment group). However, because more complex experimental designs are possible, we approach the problem more generally by allowing for unbalanced data. The key question of general interest is which genes are expressing differently over the two groups. Let $\epsilon_{j,k,l}$ be iid normal$(0, \sigma_0^2)$ measurement errors. The problem can be formulated as the ANOVA model,

$$Y_{j,k,l} = \theta_{j,0} + \mu_{j,0}I\{l = 2\} + \epsilon_{j,k,l},$$
$$j = 1, \dots, p, \quad k = 1, \dots, n_{j,l}, \quad l = 1, 2, \qquad (4)$$

where those genes that are expressing differentially correspond to indices j where $\mu_{j,0} \neq 0$. (If genes turn on, then $\mu_{j,0} > 0$; otherwise, if they turn off, then $\mu_{j,0} < 0$.)

Observe that (4) has $2p$ parameters. Many of the $\theta_{j,0}$ parameters will be non-zero, because they model the mean for the control group ($l = 1$) for a gene j. However, because our interest is only in $\mu_{j,0}$, the gene-treatment effect, it is convenient to force $\theta_{j,0}$ to be near 0 so that the effective dimension of the problem is p. A useful strategy is to replace $Y_{j,k,l}$ by the centered value,

$$Y_{j,k,l} - \overline{Y}_{j,1}, \qquad (5)$$

where $\overline{Y}_{j,l} = \sum_{k=1}^{n_{j,l}} Y_{j,k,l}/n_{j,l}$ is the mean for gene j over group l. This also reduces the correlation between the Bayesian parameters θ_j and μ_j and greatly improves the calibration of P-SVS. Section 3.2 explains this point in more detail.

Remark 2. Many extensions to (4) are possible. For example, additional covariates can be included in the model, which is useful in experiments where external data (e.g., clinical information of the individuals providing tissue samples) is collected alongside gene expressions. Thus (4) can be extended to the important class of analysis of covariance (ANCOVA) models. In another extension, genes may have different variability in addition to different mean expression levels. In such cases, (4) can be extended to allow for differing gene measurement error variances σ_j^2. Often σ_j^2 will be related to the mean gene expression. In settings where this relationship is simple (e.g., where the variance might be linear in the mean), this behavior is often handled by applying a variance-stabilizing transform. Section 6.1 studies how well this approach works. On the other hand, the relationship between σ_j^2 and the mean expression can often be complex, as is the case for the colon cancer study of Section 7. For such problems, a specialized method is developed.

Remark 3. The expressions $Y_{j,k,l}$ are typically the end products of some form of probe-level standardization across samples (i.e., chips). Standardization for the colon cancer data involves standardizing chips to a gamma distribution with a fixed shape parameter. Further gene-level preprocessing is done as discussed in Section 3.1. With so much processing of the data, it is natural to question the assumption of normality in (4). In fact, Theorem 1 in Section 3.4 shows that this assumption is not necessary because of the effect of a central limit theorem, as long as $\epsilon_{j,k,l}$ are independent with suitably behaved moments and group sizes $n_{j,l}$ are relatively large. Some empirical evidence of this effect is presented in Section 6 for Poisson data.

3.1 Linear Regression and Data Preprocessing

To harness P-SVS in the BAM method, we need to rewrite (4) as a linear regression model (1). This means that we need to write (4) using a single index, i. This is accomplished by stringing out observations (5) in order, starting with the observations for gene $j=1$, group $l=1$, followed by values for gene $j=1$, group $l=2$, followed by observations for gene $j=2$, group $l=1$, and so on. Notationally, we also need to make the following adjustments for parameters. In place of β_j we now use coefficients (θ_j,μ_j). Hierarchical prior variances for (θ_j,μ_j) are now denoted by (v_{2j-1}^2, v_{2j}^2). In a similar way, make notational changes to γ and τ^2 in (2).

The second step to using P-SVS requires a data preprocessing step that rescales the observations by the square root of the sample size divided by the mean squared error and by transforming the design matrix so that its columns each satisfy $\sum_i x_{i,j}^2 = n$. In low to moderate correlation problems (in the design matrix) this calibrates the conditional variance $(\Gamma^{-1} + X'X/\sigma^2)^{-1}$ for β in (3) to have diagonal elements nearly 0 or 1. Variances of 0 correspond to nonsignificant variables, whereas variances of 1 correspond to informative covariates. Because the conditional variance of 1 represents a good bound for the variance of a significant variable, a standard normal(0, 1) distribution can then be used to assess whether a specific regression coefficient, β_j, should be considered informative and hence included in the model (see Ishwaran and Rao 2000 for further elaboration on this point).

Let $\hat\sigma_n^2$ denote the usual unbiased estimator for σ_0^2 from (4),

$$\hat\sigma_n^2 = \frac{1}{n-2p}\sum_{j,k,l}(Y_{j,k,l} - \overline{Y}_{j,1}I\{l=1\} - \overline{Y}_{j,2}I\{l=2\})^2,$$

where $n = \sum_{j=1}^p n_j$ is the total number of observations and $n_j = n_{j,1} + n_{j,2}$ is the total sample size for gene j. To calibrate the data, replace the observations (5) with rescaled values,

$$\widetilde{Y}_{j,k,l} = (Y_{j,k,l} - \overline{Y}_{j,1}) \times \sqrt{n/\hat\sigma_n^2}.$$

The effect of this scaling is to force σ^2 to be approximately equal to n, and to rescale posterior mean values so they can be directly compared with a limiting normal distribution. We also rescale the columns of X to have second moments equal to 1. Thus after preprocessing, we can write (4) as

$$\widetilde{Y} = \widetilde{X}^T\tilde\beta_0 + \tilde\epsilon,$$

where $\widetilde{Y}$ is the vector of the n strung-out $\widetilde{Y}_{j,k,l}$ values, $\tilde\beta_0$ is the new regression coefficient under the scaling, $\tilde\epsilon$ is the new vector of measurement errors, and $\widetilde{X}$ is the $n \times (2p)$ design matrix. Here $\widetilde{X}$ is defined so that its $2j-1$ column consists of 0s everywhere except for the values $\sqrt{n/n_j}$ placed along the n_j rows corresponding to gene j, whereas column $2j$ consists of 0s everywhere except for values $\sqrt{n/n_{j,2}}$ placed along the $n_{j,2}$ rows corresponding to gene j for group $l=2$.

3.2 Ridge Regression Estimates

Because of the simple nature of the design matrix $\widetilde{X}$, we can explicitly work out the conditional distribution for β. Adjusting to the new notation, a little bit of algebra using (3) shows that

$$((\theta_j,\mu_j)^T | v_{2j-1}^2, v_{2j}^2, \sigma^2, Y) \sim \text{normal}((\hat\theta_{j,n},\hat\mu_{j,n})^T, \widehat{\Sigma}_{j,n}^{-1}),$$

where

$$\begin{pmatrix}\hat\theta_{j,n}\\ \hat\mu_{j,n}\end{pmatrix} = \frac{n\widehat{\Sigma}_{j,n}^{-1}}{\hat\sigma_n\sigma^2}\begin{pmatrix} n_j^{-1/2}\sum_{k,l}(Y_{j,k,l}-\overline{Y}_{j,1})\\ n_{j,2}^{-1/2}\sum_k(Y_{j,k,2}-\overline{Y}_{j,1})\end{pmatrix}, \qquad (6)$$

and

$$\widehat{\Sigma}_{j,n} = \begin{pmatrix} n/\sigma^2 + 1/v_{2j-1}^2 & n/\sigma^2 \times \sqrt{n_{j,2}/n_j}\\ n/\sigma^2 \times \sqrt{n_{j,2}/n_j} & n/\sigma^2 + 1/v_{2j}^2\end{pmatrix}.$$

Recall that as a consequence of the preprocessing step, σ^2 should be approximately equal to n. Also, because $\theta_{j,0}$ should be nearly 0 due to the centering (5), we should expect $v_{2j-1}^2 \approx 0$. Thus

$$\widehat{\Sigma}_{j,n}^{-1} \approx \begin{pmatrix} 0 & 0\\ 0 & v_{2j}^2/(v_{2j}^2+1)\end{pmatrix}.$$

This implies that the posterior correlation between θ_j and μ_j should be nearly 0. Thus the centering method (5) has two important roles: (a) It reduces the dimension of the parameter space and (b) it reduces correlation between parameters.

Now for genes that are expressing differently, we expect v_{2j}^2 to be large, and thus the conditional mean for μ_j is approximately

$$\hat\mu_{j,n} \approx \frac{\sqrt{n_{j,2}}}{\hat\sigma_n}(\overline{Y}_{j,2} - \overline{Y}_{j,1}), \qquad (7)$$

and the conditional variance for μ_j should be approximately $v_{2j}^2/(v_{2j}^2+1) \approx 1$. Because the conditional variance represents a lower bound for the posterior variance $V(\mu_j|Y)$, this suggests that the posterior mean $E(\mu_j|Y)$ should be compared with a normal(0, 1) distribution to test whether μ_j is 0. However, Theorem 1 of Section 3.4 shows that it is more appropriate to use the theoretical limiting variance for $\hat\mu_{j,n}$ as our lower bound, here equal to $n_j/n_{j,1}$. Thus $E(\mu_j|Y)$ should be compared with a normal$(0, n_j/n_{j,1})$ distribution to test whether $\mu_{j,0}$ is nonzero. This is the basis of the Zcut procedure.

Theorem 1 justifies this strategy more rigorously, in the meantime some evidence for this approach can be seen in Figure 2(a), which plots the posterior absolute means and posterior variances from the earlier gene simulation model. (All computations were done using the P-SVS Gibbs sampling scheme outlined in Ishwaran and Rao 2000.) In this "shrinkage plot" a thin-dashed horizontal line represents the lower variance bound of 1, and the thick-dashed horizontal line represents the theoretical variance $n_j/n_{j,1} = 2$. Notice how large values for $|E(\mu_j|Y)|$ have posterior variances near 1 that then jump up to the theoretical variance $n_j/n_{j,1} = 2$ as $|E(\mu_j|Y)|$ becomes moderate in size, finally dropping down to 0 as $|E(\mu_j|Y)|$ becomes

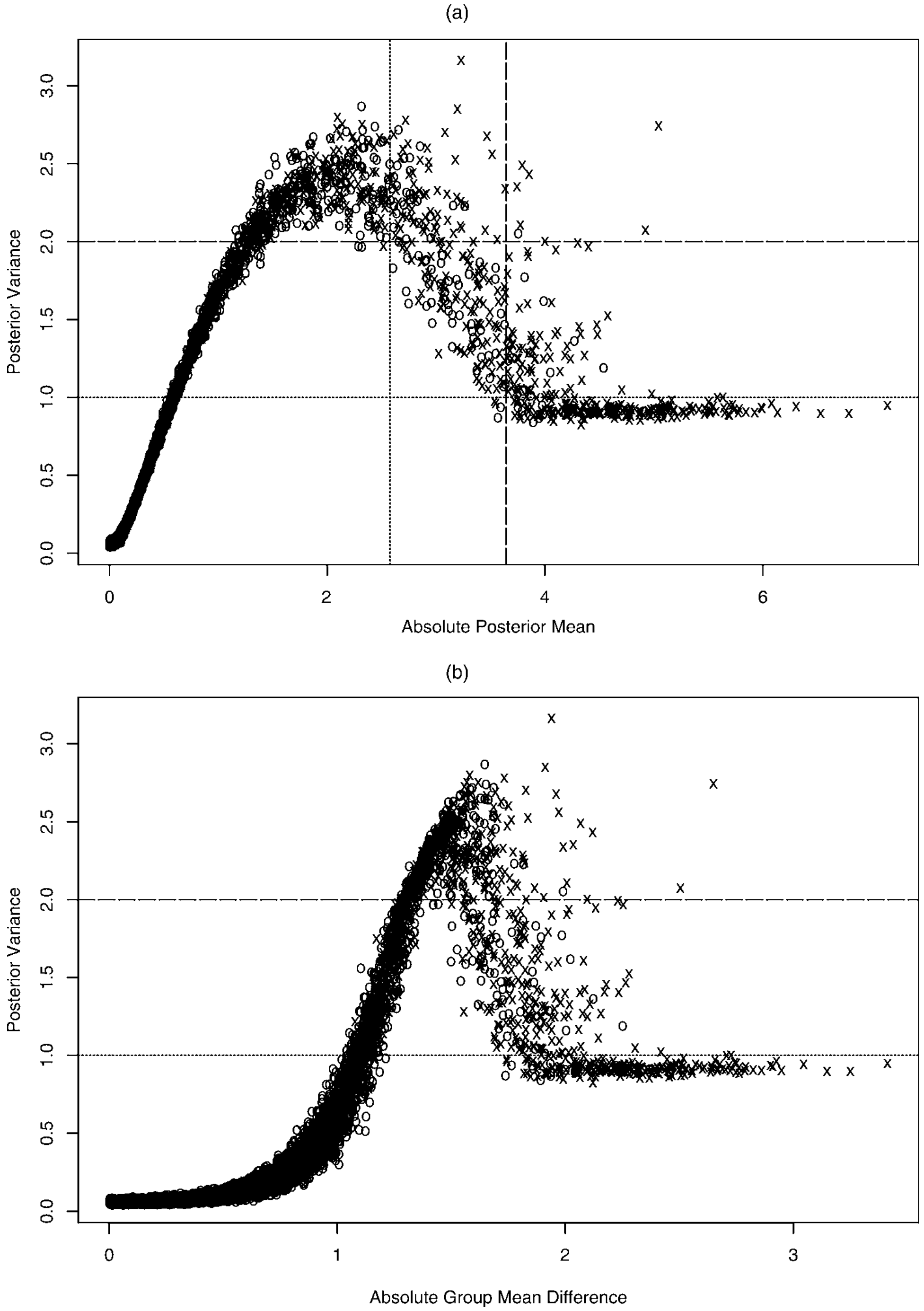

Figure 2. (a) Estimated Values for $|E(\mu_j|Y)|$ Versus var$(\mu_j|Y)$ and (b) Absolute Mean Differences $|\overline{Y}_{j,2} - \overline{Y}_{j,1}|$ Versus var$(\mu_j|Y)$ From the Gene Simulation Model of Figure 1. Circles indicate genes whose true means are $\mu_{j,0} = 0$; crosses, genes whose true means are $\mu_{j,0} \neq 0$.

small. Figure 2(b) shows that the genes with large posterior variances are those with group mean differences of intermediate size. Thus the jump is seen because BAM is shrinking the posterior means while inflating the variance for these intermediate values, making it harder to reject the null $\mu_{j,0} = 0$. It is clear that to better classify those genes with moderate values for $|E(\mu_j|Y)|$, we need to adjust the variance to equal $n_j/n_{j,1}$. As illustration, consider the two vertical lines in Figure 2(a). The thin-dashed line is the value for the 99.5th percentile from a standard normal(0, 1) distribution, whereas the thick-dashed line is the same percentile from a normal(0, 2) distribution. Observe how the adjustment to the variance helps avoid misclassifying genes.

Remark 4. It is instructive to work out the posterior conditional variance without using the centering method (5). Now we expect $\hat{\theta}_{j,n}$ to be non-zero and that v_{2j-1}^2 will be large. If $\mu_{j,0}$ is non-zero, then v_{2j}^2 will be large, and hence

$$\widehat{\Sigma}_{j,n}^{-1} \approx \begin{pmatrix} n_j/n_{j,1} & -\sqrt{n_j n_{j,2}}/n_{j,1} \\ -\sqrt{n_j n_{j,2}}/n_{j,1} & n_j/n_{j,1} \end{pmatrix}.$$

Note the resulting non-zero posterior correlation of $-\sqrt{n_{j,2}/n_j}$ between θ_j and μ_j.

3.3 The Zcut Procedure

Formally, what we call the *Zcut procedure*, or simply *Zcut*, is the following method for identifying parameters $\mu_{j,0}$ that are not 0 (and hence genes j that are expressing). Identify a gene j as differentially expressing if

$$|E(\mu_j|Y)| \geq z_{\alpha/2}\sqrt{n_j/n_{j,1}},$$

where $z_{\alpha/2}$ is the $100 \times (1-\alpha/2)$th percentile of a standard normal distribution. The value $E(\mu_j|Y)$ is obtained by averaging Gibbs sampled draws for μ_j.

Remark 5. In practice, we need an appropriate way to select an α value for Zcut. A technique that we use in Section 7 is to select α on the basis of a shrinkage plot like that in Figure 2(a). There we chose α to coincide with the vertical quantile so that most of the observations to its right will have posterior variance, $\text{var}(\mu_j|Y)$, nearly equal to 1. This removes many intermediate values that could inflate the FDR.

3.4 Robustness and Adaptiveness

We now justify the Zcut technique by way of a central limit theorem. This analysis justifies the adjustment to the variance used in Figure 2(a), identifying it with the asymptotic variance for the conditional posterior mean. These results hold even when measurement errors are not normally distributed, assuming that appropriate moment conditions are satisfied. In the following theorem, carefully note the degeneracy of the limiting distribution. This is a consequence of the centering method (5).

Theorem 1. Assume that (4) represents the true model, where $\epsilon_{j,k,l}$ are independent random variables such that $E(\epsilon_{j,k,l}) = 0$, $E(\epsilon_{j,k,l}^2) = \sigma_0^2$, and $E(|\epsilon_{j,k,l}^3|)$, $E(\epsilon_{j,k,l}^4) \leq C_0$ for some fixed constant $C_0 < \infty$. If $\sigma^2/n \to 1$ and $n_{j,1} \to \infty$ and $n_{j,2} \to \infty$ for each j such that $n_{j,2}/n_{j,1} \to r_{j,0} < \infty$, then under the null hypothesis $\mu_{j,0} = 0$, keeping (v_{2j-1}^2, v_{2j}^2) fixed,

$$(\hat{\theta}_{j,n}, \hat{\mu}_{j,n})^T \overset{\text{d}}{\rightsquigarrow} Z_j \Sigma_j^{-1}\big(\sqrt{r_{j,0}}, \sqrt{1+r_{j,0}}\big)^T, \qquad j = 1, \dots, p,$$

where

$$\Sigma_j = \begin{pmatrix} 1 + 1/v_{2j-1}^2 & \sqrt{r_{j,0}/(1+r_{j,0})} \\ \sqrt{r_{j,0}/(1+r_{j,0})} & 1 + 1/v_{2j}^2 \end{pmatrix}$$

and Z_j are independent normal(0, 1) random variables.

Most of the conditions of Theorem 1 are likely to hold in practice. Relaxing the assumption that errors are normally distributed and iid is an especially helpful feature. The condition that $\sigma^2/n \to 1$ is quite realistic and understandable because of our rescaling method. We have found that it holds quite accurately in many problems. For example, in the simulations presented in Figure 2(a), σ^2/n had a posterior mean of .96 with a posterior standard deviation of .04. The value for $r_{j,0}$ appearing in the theorem represents the limiting ratio of the group sizes. In our previous simulation, $r_{j,0} \approx n_{j,2}/n_{j,1} = 1$. If the posterior identifies gene j as differentially expressing (i.e., v_{2j}^2 is large), then Theorem 1 and an argument similar to (7) shows that $\hat{\mu}_{j,n}$ can be approximated in distribution by a normal(0, $1 + r_{j,0}$), or roughly a normal(0, 2), under the null. Notice how this matches up with the adjusted variance used in our simulation. In general, under the conditions of Theorem 1, $\hat{\mu}_{j,n}^* := \hat{\mu}_{j,n}\sqrt{n_{j,1}/n_j}$ can be approximated by a standard normal distribution under the null when v_{2j}^2 is large. If we integrate over the hyperparameters, then we get $\mu_{j,n}^* := E(\mu_j|Y)\sqrt{n_{j,1}/n_j}$, which we can think of as a "Bayes test statistic." This is the rationale for the Zcut rule defined in Section 3.3.

Theorem 1 shows that BAM is robust to underlying model assumptions and provides an explanation for the Zcut rule. Our next result, Theorem 2, translates BAM's ability to adaptively shrink coefficients into a statement about risk performance. Shrinkage of coefficients implies shrinkage of test statistics. BAM's adaptiveness allows it to shrink the Bayes test statistics $\mu_{j,n}^*$ for coefficients $\mu_{j,0} = 0$ while allowing $\mu_{j,n}^*$ from coefficients $\mu_{j,0} \neq 0$ to have values similar to frequentist Z-tests,

$$Z_{j,n} = \frac{\overline{Y}_{j,2} - \overline{Y}_{j,1}}{\hat{\sigma}_n\sqrt{1/n_{j,1} + 1/n_{j,2}}}. \tag{8}$$

This has a direct impact on performance. Because both $\mu_{j,n}^*$ and $Z_{j,n}$ are compared with a normal(0, 1) in assessing significance, Zcut will produce p values that match up closely with Z-tests for non-zero coefficients, whereas p values from the 0 coefficients will be much larger. Recall that we saw this p-value effect in Figure 1. Figure 3 illustrates the effect in terms of test statistics. This translates into power for Zcut and a low FDR.

Theorem 2 quantifies this adaptiveness in terms of risk behavior. We define some notation to explain this. Let $\hat{d}_{j,\alpha} \in \{0, 1\}$ represent the classification rule for some two-sided test at a fixed α level. A value $\hat{d}_{j,\alpha} = 1$ means that we reject the null $\mu_{j,0} = 0$; otherwise, if $\hat{d}_{j,\alpha} = 0$, then we accept the null. Let

$$R_{j,\alpha} = \Pr\{\hat{d}_{j,\alpha} = 1, \mu_{j,0} = 0\} + \Pr\{\hat{d}_{j,\alpha} = 0, \mu_{j,0} \neq 0\}.$$

This is the expected risk for gene j. The total expected risk for all genes is $R(\alpha) = \sum_{j=1}^{p} R_{j,\alpha}$. The average $R(\alpha)/p$ is the misclassification rate. Write $R_B(\alpha)$ and $R_F(\alpha)$ for the total expected risk using $\hat{\mu}_{j,n}^*$ and $Z_{j,n}$ for some fixed α value.

Theorem 2. Assume that (4) represents the true model where $\epsilon_{j,k,l}$ are iid normal(0, σ_0^2) variables. Suppose that $\sigma^2 = n$. Then for each $0 < \delta < 1/2$ there exist values (v_{2j-1}^2, v_{2j}^2) for $j = 1, \dots, p$ such that $R_B(\alpha) < R_F(\alpha)$ for each $\alpha \in [\delta, 1-\delta]$.

Theorem 2 shows that over a wide range of α values, with suitably selected values of (v_{2j-1}^2, v_{2j}^2), the expected total risk for Zcut will be less than that for the classification rule derived from using Z-tests. The conditions of the theorem are fairly reasonable. The restriction to normally distributed errors is made for convenience, because a central limit theorem as in Theorem 1 should apply in practice. Section 6 provides several simulations verifying Zcut's better risk performance over the use of Z-tests.

4. GENERATING THE NULL: FDRMIX

Values estimated from BAM can also be used in a hybrid BH procedure as a method for controlling the FDR. We call this new model selection procedure *FDRmix*. Like the Zcut procedure, FDRmix selects models based on posterior values for μ_j. In

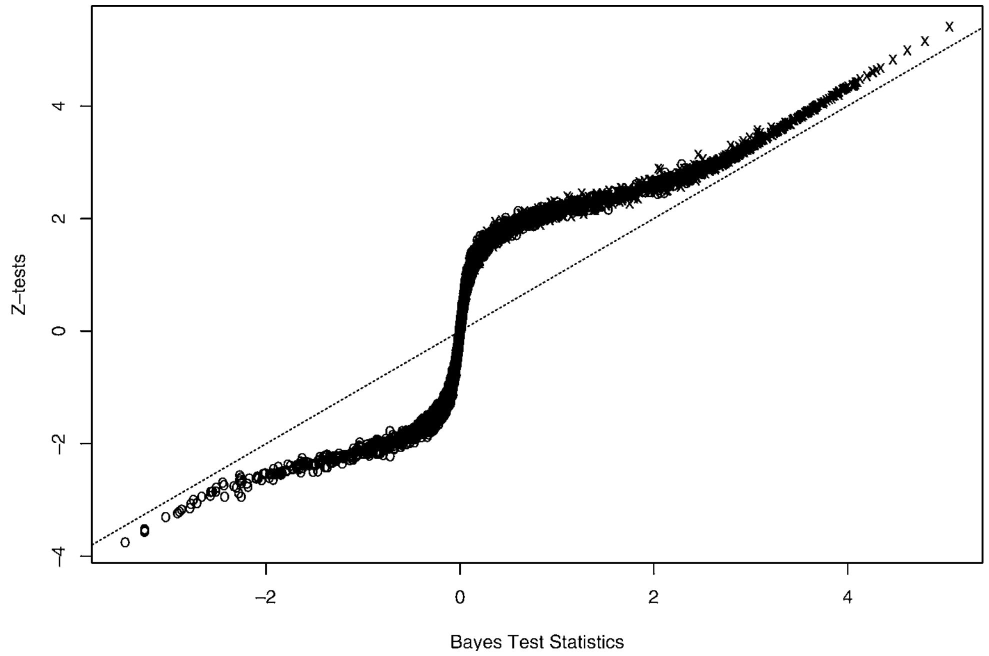

Figure 3. BAM Test Statistics $\mu^*_{j,n}$ Versus $Z_{j,n}$ From Simulations in Figure 1. Expressed genes are represented by crosses; nonexpressed genes, by circles. (Nonexpressed genes are the values mostly near 0 on the x-axis that have been shrunken by BAM.)

this approach, we use $T_j = E(\mu_j|Y)$, the posterior mean value for μ_j, as the test statistic in selecting models.

To implement FDRmix, we need to derive $F_0(dT_j)$, the conditional density for T_j under the null hypothesis $\mu_{j,0} = 0$. Although an exact derivation is infeasible, an accurate and simple approximation for F_0 can be derived by considering $\hat{\mu}_{j,n}$, given earlier in (6). Suppose that the null is true. Although the posterior should identify gene j as having a mean of 0, there will still be a positive posterior probability of a large v^2_{2j} and a resulting misclassification. Equation (7) identifies T_j under this scenario. Suppose that the data are balanced with fixed group sizes so that $n_{j,1} = N_1$ and $n_{j,2} = N_2$. Then, under the null, conditioning on the event $A_j = \{v^2_{2j}$ is large$\}$,

$$\begin{aligned}(T_j|A_j, \mu_{j,0} = 0) &\approx \frac{\sqrt{N_2}}{\hat{\sigma}_n}(\overline{Y}_{j,2} - \overline{Y}_{j,1})\\ &\approx \text{normal}(0, (N_1 + N_2)/N_1).\end{aligned}$$

On the other hand, if the posterior correctly identifies that gene j is not expressing (i.e., that $A^c_j = \{v^2_{2j} \approx 0\}$), then (6) suggests that

$$(T_j|A^c_j, \mu_{j,0} = 0) \approx \frac{\sqrt{N_2}}{\hat{\sigma}_n(1 + 1/v^2_{2j})}(\overline{Y}_{j,2} - \overline{Y}_{j,1})$$

for some small value of v^2_{2j}. Under the null, this also has a normal distribution with mean 0, but unlike in the previous case, the theoretical variance here is not so clear.

These arguments suggest that we can approximate the null distribution of T_j using the two-point normal mixture,

$$F_0(dT_j) \approx (1 - \Pi_0)\phi(T_j|V_1) + \Pi_0\phi(T_j|V_2),$$

where $\phi(\cdot|V)$ denotes a normal density with mean 0 and variance V. We anticipate that $V_2 = (N_1 + N_2)/N_1$ but the values for V_1 and $\Pi_0 = \Pr\{A_j|\mu_{j,0} = 0\}$ are unknown. All of these values, however, can be easily estimated by fitting a two-point normal mixture to simulated data. Thus to estimate V_1, V_2, and Π_0, we simulate data from the model (4), where $\mu_{j,0} = 0$ for all $j = 1, \dots, p$. (Typically we would choose p to be some large number; here we used 25,000.) Notice that this simulation requires knowing only the sample sizes N_1 and N_2 and the value for σ^2_0, which can be estimated accurately from the original data. We then run the BAM procedure on the simulated data and fit a two-point mixture model to the averaged values for μ_j collected from the Gibbs output. The results from fitting the mixture can now be used to derive p values, which are then analyzed in the usual way by the BH method to determine which hypotheses to reject. To compute the p values, suppose that T^o_j is the estimated value for $E(\mu_j|Y)$ from the original (nonsimulated) data. Then its p value, P_j, can be approximated by

$$\begin{aligned}P_j &= 2\Pr\{|T^o_j| < T_j|\mu_{j,0} = 0\}\\ &\approx 2(1 - \Pi_0\Phi(|T^o_j|/\sqrt{V_2}) - (1 - \Pi_0)\Phi(|T^o_j|/\sqrt{V_1})),\end{aligned}$$

where $\Phi(\cdot)$ denotes a standard normal cdf.

5. COMPARISON PROCEDURES

We tested BAM against several different model selection procedures, including what we call “Bayes exch,” “Z-test,” “Bonf,” and “BH.” Here we give brief descriptions of these methods.

Z-test. Here genes are identified by the Z-test statistics defined earlier in (8). The Z-test procedure identifies gene j as expressing if $|Z_{j,n}| \geq z_{\alpha/2}$.

Bonf. The Bonf procedure is a Bonferroni correction to Z-test. Thus gene j is identified as expressing if $|Z_{j,n}| \geq z_{\alpha/(2p)}$.

BH. We also applied the BH procedure. The p values used were based on the test statistics $Z_{j,n}$ under a two-sided test. Thus if $P_j = 2\ \Pr\{\text{normal}(0,1) > |Z_{j,n}|\}$, then gene j is considered expressing if $P_j \le P_{(k)}$, where $P_{(k)}$ is the kth-ordered p value, where $k = \max\{j : P_{(j)} \le j\alpha/p\}$ and $\alpha > 0$ is some prechosen target FDR. Although the original BH procedure was designed to control the expected FDR assuming independent null hypotheses, Benajamini and Yekutieli (2001) showed that the method applies under certain types of dependencies. Corollary 3.3 of Benajamini and Yekutieli (2001) showed that the method computed from dependent Z-tests such as $Z_{j,n}$ will control the expected FDR if applied to those $Z_{j,n}$ for which the null is true. Thus when many nulls are true, this should approximately control the expected FDR.

Bayes exch. We also implemented a simple Bayesian exchangeable model similar in nature to that of Gelman and Tuerlinckx (2000). To model (4), we replace the data $Y_{j,k,l}$ with sufficient statistics $\overline{Y}_{j,2} - \overline{Y}_{j,1}$. Conditional on $\hat{\sigma}_n^2$ this should have a normal$(\mu_j, \hat{\sigma}_n^2(1/n_{j,1} + 1/n_{j,2}))$ distribution. Thus to identify genes, we used the (empirical) hierarchical model

$$(\overline{Y}_{j,2} - \overline{Y}_{j,1} | \mu_j, \hat{\sigma}_n^2) \sim \text{normal}(\mu_j, \hat{\sigma}_n^2(1/n_{j,1} + 1/n_{j,2})),$$

$$(\mu_j | \tau_0^2) \sim \text{normal}(0, \tau_0^2),$$

$$(\tau_0^{-2} | t_1, t_2) \sim \text{gamma}(t_1, t_2),$$

where $t_1 = t_2 = .0001$ was selected to ensure a noninformative prior for τ_0^2. This extends the models considered by Gelman and Tuerlinckx (2000) by allowing for a hyperprior on τ_0^2.

Observe that given the data Y, the hyperparameter τ_0^2 and the estimate $\hat{\sigma}_n^2$,

$$\begin{aligned}(\mu_j | Y, \hat{\sigma}_n^2, \tau_0^2) \sim \text{normal}(&\tau_0^2(\overline{Y}_{j,2} - \overline{Y}_{j,1})/(S_{j,n}^2 + \tau_0^2),\\ &S_{j,n}^2 \tau_0^2/(S_{j,n}^2 + \tau_0^2)),\end{aligned}$$

where $S_{j,n}^2 = \hat{\sigma}_n^2(1/n_{j,1} + 1/n_{j,2})$. Thus a reasonable procedure identifies gene j as expressing if

$$|\overline{Y}_{j,2} - \overline{Y}_{j,1}| \ge z_{\alpha/2} S_{j,n} \sqrt{1 + S_{j,n}^2/\hat{\tau}_0^2}, \qquad (9)$$

where $\hat{\tau}_0^2$ is some estimate for τ_0^2. In all examples, we took $\hat{\tau}_0^2 = E(\tau_0^2 | Y, \hat{\sigma}_n^2)$.

Remark 6. Notice that when $\hat{\tau}_0^2 \to \infty$, the limit of the test (9) corresponds to a standard Z-test. However, whenever $\hat{\tau}_0^2 < \infty$, the Bayesian test (9) will always be more conservative (see Gelman and Tuerlinckx 2000 for further discussion).

6. SIMULATIONS

To assess the procedures, we tested them on simulated data from the ANOVA model (4). Means for genes were set to have a group mean separation of $\mu_{j,0} = 1.5$ for 10% of parameters and $\mu_{j,0} = 0$ for 90% of parameters. This corresponds to a model in which 10% of genes are expressing and represents a fairly realistic scenario for microarray data. For convenience, we set $\theta_{j,0} = 0$ for all j. Thus our gene simulation model was

$$Y_{j,k,l} = \mu_{j,0} I\{l = 2\} + \epsilon_{j,k,l},$$

$$j = 1, \ldots, p, \quad k = 1, \ldots, n_{j,l}, \quad l = 1, 2,$$

where $\epsilon_{j,k,l}$ were taken to be iid normal$(0, \sigma_0^2)$ variables. For the variance, we set $\sigma_0^2 = 1$. Group sizes were fixed at $n_{j,1} = n_{j,2} = 5$, and the number of genes was fixed at $p = 25{,}000$. (The simulation reported in Table 1 and Figure 1 used this same configuration, but with $p = 10{,}000$.)

All model estimators reported were based on a range of preset α levels. For Zcut, Bayes exch, Z-test, and Bonf, a preset α value meant that the corresponding test was computed based on the appropriate quantile for a standard normal; for example, Z-test was computed using $z_{\alpha/2}$ as a cutoff, whereas Bonf used $z_{\alpha/(2p)}$. For FDRmix and BH, the α value used was the target FDR. To be able to fairly compare the procedures under the same α values, we recorded the total number of misclassified observations, TotalMiss, discussed in Section 1. Also recorded was the FDR, FNR, and the type I and type II error rates.

The results of the simulations are reported in Table 2 and Figure 4. From these, we make the following general observations:

1. Z-test has the best total risk performance for small α values, but its good risk behavior is only local. For example, once α increases, the value for TotalMiss increases rapidly, and procedures like Zcut and FDRmix quickly have superior risk performance. Furthermore, if α becomes very small, then its TotalMiss values also shoot up. Thus, trying to find the small region of α values where Z-test does well is a tricky proposition. No theory exists to select this region, which makes Z-test unreliable. Overall, Zcut and FDRmix have the best total risk behavior. For small to moderate α values, Zcut is better, whereas for large α FDRmix is better. BH has poorer performance for small to moderate α, approaching the same performance as Zcut and FDRmix only for α as large as .2. This confirms BH's inability to simultaneously optimize the FDR and FNR. Bayes exch tracks BH in terms of total risk over small to moderately large α.
2. FDRmix does not achieve target α values, but does further reduce the FDR for Zcut as designed. The FDR for BH is better, but is generally smaller than its target value. When α becomes small, its FDR drops to 0 and it becomes too conservative.
3. Of all the procedures, Zcut has type II error rates closest to those observed for Z-test.

Table 2. Gene Simulation Model $p = 25{,}000$ With 10% of Genes Expressing

	Detected	*TotalMiss*	*FDR*	*FNR*	*Type I*	*Type II*
Zcut	1,286	1,700	.189	.061	.011	.583
	822	1,824	.088	.072	.003	.700
	281	2,231	.021	.090	0	.890
FDRmix	1,148	1,724	.162	.064	.008	.615
	719	1,885	.072	.076	.002	.733
	72	2,430	.014	.097	0	.972
Bayes exch	498	2,048	.046	.083	.001	.810
	78	2,424	.013	.097	0	.969
	3	2,497	0	.099	0	.999
Z-test	2,808	2,006	.412	.038	.051	.340
	1,245	1,709	.182	.062	.010	.593
	453	2,081	.038	.084	.001	.826
Bonf	18	2,482	0	.099	0	.993
	10	2,490	0	.099	0	.996
	2	2,498	0	.099	0	.999
BH	415	2,115	.036	.085	.001	.840
	116	2,388	.017	.096	0	.954
	11	2,489	0	.099	0	.996

NOTE: Estimators are based on α levels of .05, .01, and .001.

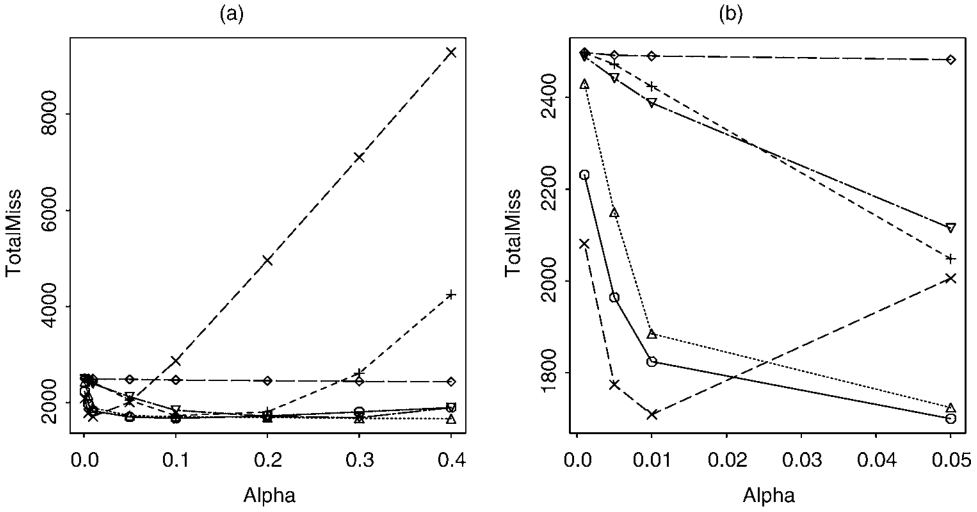

Figure 4. (a) Total Number of Misclassified Observations From Gene Simulation Model p = 25,000. (b) Close-up View For Smaller α Values. (○ —— Zcut, △..... FDRmix, + - - - Bayes exch, × – – Z-test, ◇ — — Bonf, ▽ — - BH).

4. The worst procedure overall is Bonf, whose TotalMiss plot is flat. Alternatives to Bonferroni corrections have been attempted in the microarray literature (see, e.g., Callow, Dudoit, Gong, Speed, and Rubin 2000). These often involve nonparametric stepdown adjustments to raw p values using permutation distributions of the test statistics. Modest power gains have been seen, but coarseness of the permutation distributions limits the usefulness of these approaches to situations in which a large enough number of samples are available. To illustrate, we applied the adjusted p-value method of Holm (1979) and found that with α equal to .05, .01, and .001, we picked up 18, 10, and 2 significant genes. This performance mirrors that of Bonf. Because step-down adjustment methods also attempt to control familywise error rates, it is not surprising that we found them to be quite conservative.

Remark 7. One referee made the interesting remark that although Figure 4 shows that Z-test can be quite volatile near its optimal α value, it seems to suggest that Z-test nonetheless outperforms Zcut for smaller α values. Thus Z-test's volatility may not be a problem as long as one adopts the strategy of choosing a small α value. In fact, this can be a bad strategy, because the TotalMiss values for Z-test can sometimes rise rapidly even at near-0 values for α. To illustrate this behavior, we reduced the signal-to-noise ratio of the previous simulation by setting $\mu_{j,0} = .5$ for the 10% of genes that were expressing. All other values of the simulation were kept the same. Figure 5(a) records the TotalMiss values for Z-test and Zcut from the analysis. The plot clearly shows that Zcut has superior risk performance over Z-test even at near-0 α values. Note also how the TotalMiss values for Z-test increase rapidly due to overfitting, whereas the risk performance for Zcut stabilizes. The better risk performance of Zcut is because of its adaptive shrinkage ability as predicted by Theorem 2. Figure 5(b) shows the extent to which BAM is able to shrink test statistics compared with Z-test. It

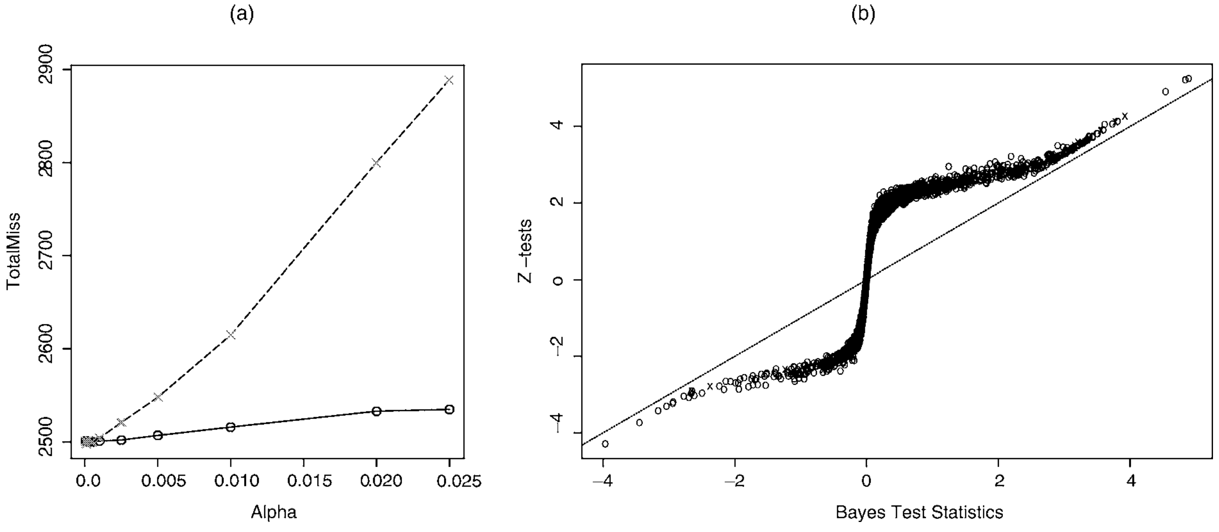

Figure 5. (a) Total Number of Misclassified Observations From Gene Simulation Model Used in Figure 4 but With Smaller Signal-to-Noise Ratio, and (b) BAM Test Statistics Against Z-Tests Similar to Figure 3. (○ —— Zcut, × - - - Z-test).

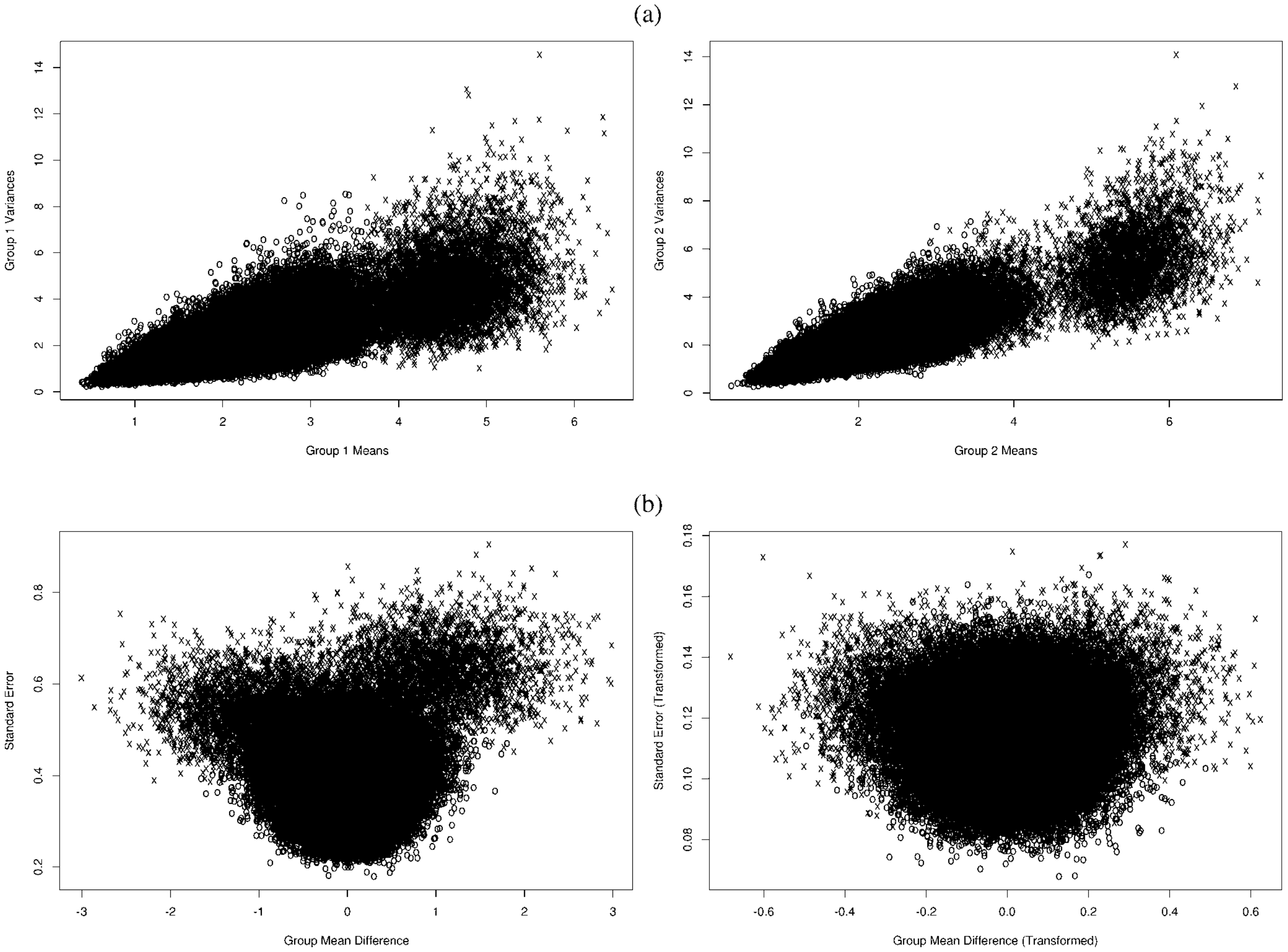

Figure 6. (a) Gene Mean Values $\overline{Y}_{j,l}$ Versus Gene Sample Variances $\hat{\sigma}^2_{j,l}$ (Left, Group $l = 1$; Right, Group $l = 2$) and (b) Group Mean Differences $\overline{Y}_{j,2} - \overline{Y}_{j,1}$ Versus Standard Errors (Left, Untransformed Data; Right, Computed From the Variance-Stabilized Data). Expressing genes are identified by crosses; nonexpressing genes, by circles.

is instructive to compare this with Figure 3 to see how much shrinkage is actually occurring.

6.1 Unequal Variances: Poisson Gene Simulation Model

Untransformed gene expression data often fail to satisfy the assumptions of an equal variance model. An often-seen pattern is that the variance for a gene expression is related to its mean expression. Often the relationship is linear, but sometimes it can be quite complex, such as in the case of colon cancer (Sec. 7). To deal with these problems, it is generally recommended that the data be transformed through a variance-stabilizing transformation (Speed 2000).

To study how well the methods perform under the scenario in which variances are proportional to the mean, and also to see how robust they are to the assumption of normality, we simulated data from the Poisson model,

$$Y_{j,k,l} = \xi_{j,k,l} + \epsilon_{j,k,l},$$
$$j = 1, \ldots, p, \quad k = 1, \ldots, n_{j,l}, \quad l = 1, 2,$$

where $\xi_{j,k,l}$ are independent Poisson($\mu_{j,l}$) variables independent of $\epsilon_{j,k,l}$, the normal($0, \sigma_0^2$) measurement errors. We set $\sigma_0^2 = .01$ to a small value. This slightly smooths the data, although at the same time the variance for a gene expression is proportional to its mean [Fig. 6(a)]. For gene group sizes we used $n_{j,1} = 21$ and $n_{j,2} = 34$, and we set the number of genes at $p = 60{,}000$. These sample sizes were selected to match those of the colon cancer dataset. We set 90% of the genes to have equal means over both groups, so that for these genes $\mu_{j,l} = \mu_j$, where μ_j was drawn randomly from a uniform distribution on (1, 3). For the remaining 10% of the genes (the expressors), we sampled the group 1 means, $\mu_{j,1}$, independently from a uniform distribution on (4, 5), and then set the group 2 means $\mu_{j,2}$ to this value, incrementing it by either $+1$ or -1 randomly with equal probability. This corresponds to genes for the treatment turning on or off.

The results of the Poisson simulation are reported in Figure 7, which plots the total number of misclassified observations as α is varied. Figure 7(a) is from the untransformed data, whereas Figure 7(b) shows results from the data under a square-root variance-stabilizing transformation. We draw the following conclusions:

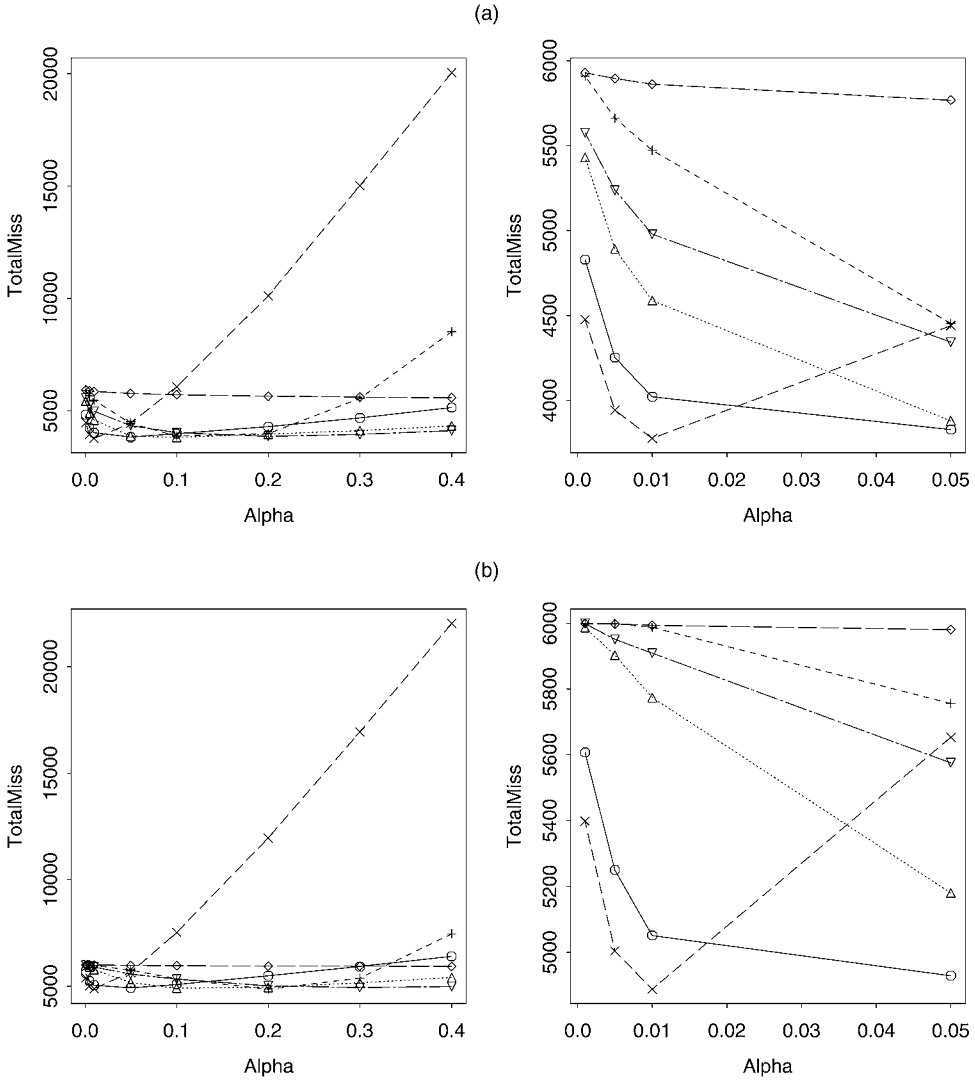

Figure 7. (a) Total Number of Misclassified Observations From the Poisson Gene Simulation Model and (b) Total Number of Misclassified Observations From the Variance-stabilized Data. (○ —— Zcut, △...... FDRmix, + - - - Bayes Exch, × – – Z-test, ◇ —— Bonf, ▽ — - BH).

1. The variance-stabilizing transform reduces the size of models across all estimators, with fewer genes being identified as expressing. In all cases, the value for TotalMiss increases under the transformation.
2. All estimators generally exhibit the same patterns of behavior as in the previous simulation.

The better performance of the methods on the untransformed data can be explained by Figure 6(b). The left side plots, for the untransformed data, the group mean difference, $\overline{Y}_{j,2} - \overline{Y}_{j,1}$, for each gene versus the corresponding standard error, $(\hat{\sigma}^2_{j,1}/n_{j,1} + \hat{\sigma}^2_{j,2}/n_{j,2})^{1/2}$, where

$$\hat{\sigma}^2_{j,l} = \frac{1}{n_{j,l} - 1} \sum_{k=1}^{n_{j,l}} (Y_{j,k,l} - \overline{Y}_{j,l})^2$$

is the sample variance for gene j over group l. We see that even though the standard error is not constant over the group mean differences (compare this with the square-root–transformed plot on the right side), its value still remains fairly constant, becoming elevated only for large group mean differences, where it still remains relatively small in magnitude compared to the mean effect. Thus, even though in the untransformed data the gene sample variances are proportional to their means [Fig. 6(a)], once the sample size is taken into account when computing the standard error, the effect of unequal variances gets washed out. It seems better to simply not transform the data, because a side effect of this is that means are compressed, leading to smaller models.

6.2 Dependence: Correlations Between Genes

Our simulations up to this point have assumed that genes are independent, but in practice it is more realistic to expect gene expressions to be correlated. A gene's expression across different samples can be safely assumed to be independent in experimental designs in which for a fixed j, the values of $Y_{j,k,l}$ represent expressions measured over different tissue types, l, for different samples, k. Within a given sample k, however, different genes can cluster into small groups—often as a manifestation of gene proximities along biological pathways. This has been termed "clumpy dependence" (Storey and Tibshirani 2001) and is the most likely scenario of dependence for these type of experimental designs. To study how the procedures perform under this form of dependence, we simulated data according to the model

$$Y_{j,k,l} = \mu_{j,0} I\{l=2\} + \zeta_{m_j,k,l} + \epsilon_{j,k,l},$$
$$j = 1,\dots,p, \quad k = 1,\dots,n_{j,l}, \quad l = 1,2,$$

where $\epsilon_{j,k,l}$ are iid normal$(0,\sigma_0^2)$ measurement errors, independent of $\zeta_{m_j,k,l}$. We used the same dimensions and group sizes as in our earlier set of simulations: $p = 25{,}000$ and $n_{j,1} = n_{j,2} = 5$. Variables $\zeta_{m_j,k,l}$ were introduced to induce dependence among genes into small "blocks." Here $m_j = [j/B]$, where $B = 50$ and $[j/50]$ represents the first integer greater than or equal to $j/50$. Thus, for each fixed k and l, there were 500 variables $\zeta_{1,k,l},\dots,\zeta_{500,k,l}$. These each induce a block of B-dependent observations. For example, "block m" comprises those $Y_{j,k,l}$ where $[j/B] = m$, which are dependent because they share the common value $\zeta_{m,k,l}$. Variables across different values for k and l were assumed to be independent; that is, $\zeta_{j,k,l}$ was independent of $\zeta_{j,k',l'}$ if $k \neq k'$ or $l \neq l'$. This ensured that a gene's expression was independent across samples.

We set the first 2,500 (10%) of genes to have non-zero $\mu_{j,0}$ coefficients and the remaining 90% to have 0 coefficients. This ensured that blocks were composed entirely of expressing genes or entirely of nonexpressing genes. All $\zeta_{m_j,k,l}$ variables were drawn from a normal(0, η_0^2) distribution. We took $\eta_0^2 = 19$ and $\sigma_0^2 = 1$, which induced a correlation of $\rho_0 = .95$ between observations within the same block. So that the signal-to-noise ratio was similar to that in our earlier simulation, we took $\mu_{j,0} = 1.5/\sqrt{1-\rho_0}$ for the expressing genes. Table 3 presents the results in a style similar to that of Table 2 for direct comparison. The values reported in Table 3 were obtained by averaging over 100 independent simulations. For brevity, only the results for Zcut, FDRmix, Zcut, and BH are reported. The values for TotalMiss and number of genes detected were rounded to the nearest integer.

Comparing Table 3 with Table 2, we make the following observations:

1. The FDR procedures, FDRmix and BH, start to break down as α decreases. The number of detected genes drops rapidly, and the FDR becomes essentially 0. FDR procedures will have high variability in dependence scenarios and will be unreliable. Benajamini and Yekutieli (2001) and Storey and Tibshirani (2001) have presented techniques for correcting FDR methods under dependence.
2. On the other hand, Zcut and Z-test have performance measurements similar to those given in Table 2. Overall, clumpy dependence seems to have a minimal effect on these. One explanation for this is that these procedures classify genes based on test statistics that are affected by dependence only through $\hat\sigma_n^2$. As long as block sizes B are relatively small compared with p (a likely scenario with microarray data), the effect of dependence will be marginal, because $\hat\sigma_n^2$ will be a robust estimator for the variance, $\sigma_0^2 + \eta_0^2$.

Table 3. Dependent Gene Simulation Model

	Detected	*TotalMiss*	*FDR*	*FNR*	*Type I*	*Type II*
Zcut	1,299	1,702	.191	.061	.011	.581
	807	1,868	.109	.074	.004	.712
	286	2,230	.029	.089	0	.889
FDRmix	1,146	1,738	.166	.065	.008	.618
	686	1,939	.094	.077	.003	.751
	33	2,467	0	.099	0	.987
Z-test	2,762	1,979	.404	.039	.049	.343
	1,269	1,698	.184	.062	.010	.586
	473	2,078	.055	.084	.001	.821
BH	403	2,129	.029	.086	.001	.845
	66	2,435	0	.098	0	.974
	4	2,496	0	.099	0	.999

NOTE: Estimators are based on α levels of .05, .01, and .001.

7. METASTATIC SIGNATURES FOR COLON CANCER

As a practical illustration of the BAM method, we now return to the issue of detection of a metastatic gene expression signature for colon cancer. This problem is of significant medical importance. Colorectal cancer is the second-leading cause of cancer mortality in the adult American population, accounting for 140,000 new cases and 60,000 deaths annually (Cohen, Minsky, and Schilsky 1997). The current clinical classification of colon cancer is based on the anatomic extent of the disease at the time of resection. It is known that this classification scheme is highly imperfect in reflecting the actual underlying molecular determinants of colon cancer behavior. For instance, upward of 20% of patients whose cancers metastasize to the liver are not given life-saving adjuvant chemotherapy based on the current clinical staging system. Thus there is an important need for the identification of a molecular signature that will identify tumors that metastasize. In addition, this signature will no doubt suggest new targets for the development of novel therapies.

The gene expression data that we consider come from the Ireland Cancer Center at Case Western Reserve University. The Center has collected a large database of gene arrays on liver metastasized colon cancers (METS), modified Duke's B1 and B2 survivors (B survivors), as expressed by the Astler–Coller–Duke's staging system (Cohen et al. 1997), and normal colon tissue samples. The B survivors represent an intermediate stage of tumor development, with metastasis to the liver seen as the disease worsens. As of February 2002, a total of 76 samples were available for analysis, made up of 34 METS, 21 B survivors, and 21 normals. The gene arrays used in compiling the database were EOS Biotechnology 59K-on-one chips, which are custom-designed Affymetrix derivations that use a smaller subset of 8 more-sensitive probes for each gene than do the original Affy chips, which use 20 probes per gene. This is based on a proprietary algorithm developed at EOS Biotechnology. The obvious advantage of such a system is the ability to assay many more genes on each chip. In fact, yields of up to 60,000 pieces of genetic information can be processed on each chip. Although some of this information might be duplicate or overlapping in nature, we treat all of the genes as independent, realizing that there is still an open debate as to the total number of actual genes in the human genome (see, e.g., Ewing and Green 2000; Crollius et al. 2000).

Our analysis was based on the data for the liver METS and B-survivor tissue samples. (We excluded the normal tissue samples, because our interest is understanding cancer evolving from

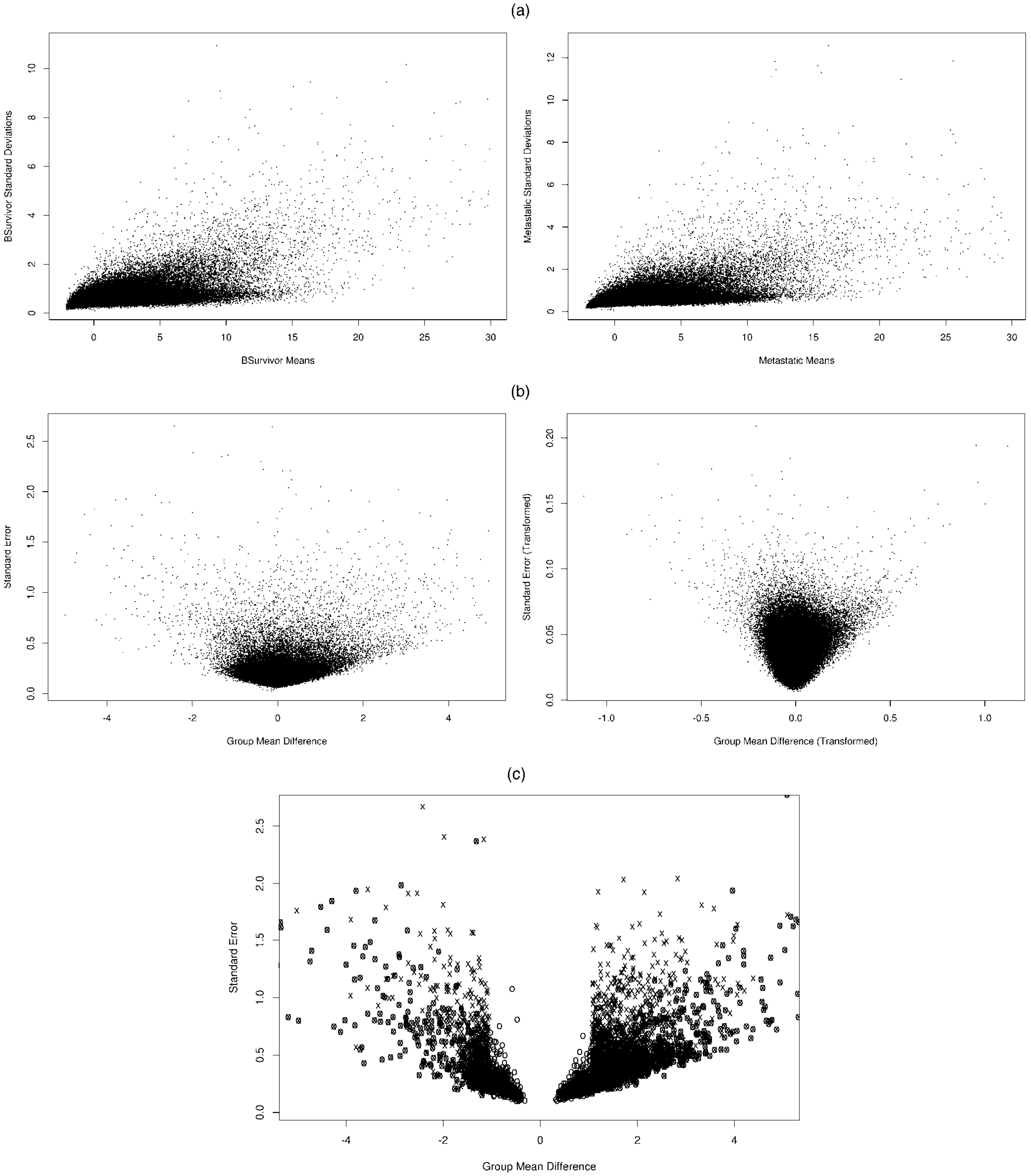

Figure 8. (a) Gene Mean Values $\overline{Y}_{j,l}$ Versus Gene Standard Deviations $\hat{\sigma}_{j,l}$ From the Colon Cancer Data. Only mean values less than 30 are plotted, to help zoom-in the plot. (b) Group mean differences $\overline{Y}_{j,2} - \overline{Y}_{j,1}$ versus standard errors. The left side was computed from untransformed data; the right side, from the log-transformed data. For the untransformed data, only mean differences less than 5 in absolute value are plotted, to help zoom-in the plot. (c) Genes found to be significant using Zcut with $\alpha = .0005$. Crosses represent genes detected from untransformed data; circles, genes detected from log-transformed data.

the intermediate B-survivor stage.) Using the earlier notation, group sizes were $n_{j,1} = 21$ for B survivors (the control group) and $n_{j,2} = 34$ for the liver METS (the treatment group). In total, there were $n_j = 55$ samples for each of $p = 59,341$ probe sets. Figure 8 plots different summary values for the data. Consider Figure 8(a), which plots the mean for a specific gene against its standard deviation to identify any potential mean–variance relationship. The figure reveals a pattern that at first glance appears to be linear. However, careful inspection shows that although standard deviations are generally increasing with the mean expression for a gene, a broad strip of values where standard deviations are essentially constant can also be seen. Thus the pattern here is one of overdispersed variances.

This raises the question of how to handle the unequal variances. One approach might be to try different transformations in the hope of stabilizing the variance. Consider the right side of Figure 8(b), which is the result of the log-transformation $\log(Y_{j,k,l} + 1 + \Delta)$, where $\Delta = |\min\{Y_{j,k,l}\}|$. (We added the constant $1 + \Delta$ to ensure that gene expressions were strictly larger than 1, so that logs for small values were not unduly inflated.) As we can see, the transformation has helped stabilize the standard error when compared with the mean differences between the two tissue groups. As discussed in Section 6.1, the stability of the standard error is the key criterion for assessing validity of inference based on an equal variance model. However, as also discussed, the gain in stabilizing standard errors by a transformation can sometimes be offset by the manner in which mean values are nonlinearly transformed. Evidence that this is happening for these data is given in Figure 8(c), which plots the genes detected by Zcut on both the untransformed and log-transformed data using the two-group ANOVA model of Section 3 (Zcut based on an $\alpha = .0005$ cutoff criteria). What we see is that the log-transform appears to be encouraging Zcut to pick genes with very small mean differences, and thus the effect of the transform is to overly enhance the signal-to-noise ratio for these genes. Such problems with the log-transform have been noticed elsewhere; for example, Rocke and Durbin (2001) and Durbin, Hardin, Hawkins, and Rocke (2002) argued against the use of log-transforms in the analysis of gene expression data.

Nonetheless, some form of variance stabilization is needed here, because the left side of Figure 8(b) reveals that the standard errors for the untransformed data are quite variable, certainly more than what we saw in the Poisson simulations of Section 6.1. Global transformations to the data were unsuccessful (we tried transformations other than the log and found them equally undesirable), and thus we relied on the classical method of weighted regression to deal with heteroscedasticity. In this approach, we grouped genes into $C = 2$ clusters depending on whether the standard error for a gene was less than or greater than the 99th percentile for standard errors. This allows us to treat the small group of genes with very high variability differently from the remaining genes. For each group, we then rescaled gene expressions $Y_{j,k,l}$ by dividing by the square root of the group mean squared error (the unbiased estimator for the variance). We then applied our methods as usual to the newly rescaled data. This method can be applied for any number of clusters C, although we generally recommend that as little transformation as possible be done to the data when trying to reduce variability. (See App. A for technical details.)

Figure 9 gives the shrinkage plot of the posterior absolute means and posterior variances for μ_j from the Gibbs output using the weighted regression approach. Once again a thin-dashed horizontal line represents a variance of 1, whereas the thick-dashed horizontal line represents the asymptotic value $n_j/n_{j,1} = 55/21$. The two vertical lines in the figure represent the $100 \times (1 - \alpha/2)$th percentile from the appropriate normal distributions for $\alpha = .0005$. The thin-dashed line is the uncorrected normal(0, 1) distribution, whereas the thick-dashed line is from the adjusted normal$(0, n_j/n_{j,1})$ distribution. The value for $\alpha = .0005$, which we used for the cutoff value in the analysis, was chosen here by eyeballing Figure 9 and picking

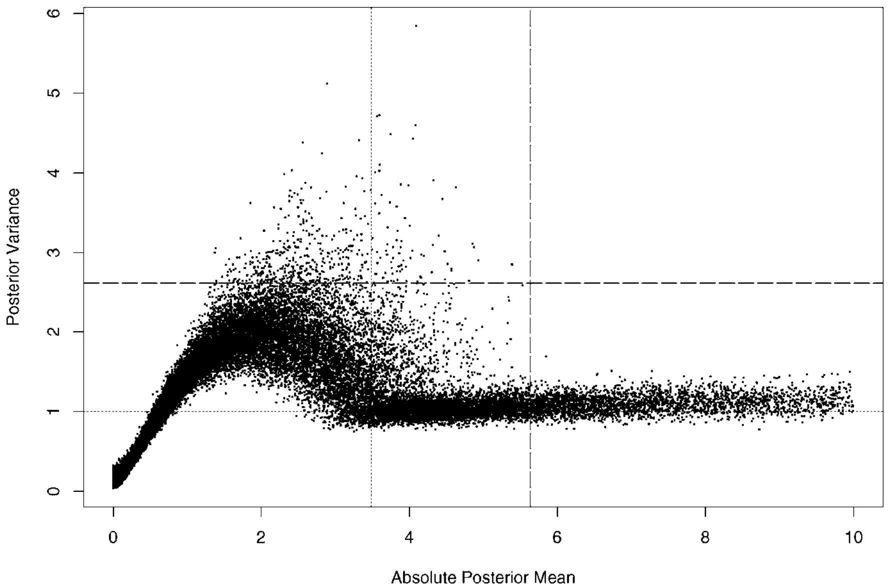

Figure 9. Estimated Values for $|E(\mu_j|Y)|$ Versus $\mathrm{var}(\mu_j|Y)$ Using Weighted Regression. Only values for $|E(\mu_j|Y)|$ less than 10 are plotted, to help zoom-in the plot.

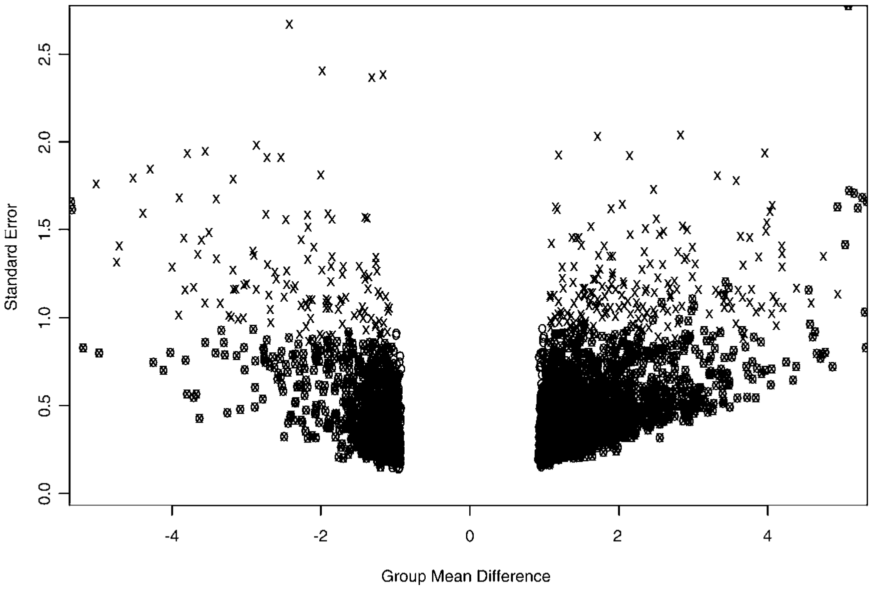

Figure 10. Genes Found to be Significant Using Zcut With $\alpha = .0005$. Crosses represent genes detected from original data; circles, genes detected using the weighted regression approach. The group mean differences and standard errors plotted are computed from the original data similar to Figure 8(c). Only mean differences less than 5 in absolute value are plotted, to help zoom-in the plot.

the vertical line that makes almost all observations to its right have posterior variance of roughly 1. This removes intermediate values that could inflate the FDR. We used the same technique to select the value of $\alpha = .0005$ in Figure 8(c).

Figure 10 plots the genes identified by Zcut from the original data and for the new data formed by the weighted regression approach. Observe how the weighting allows Zcut to pick up moderate to high expression genes, as we desire, while allowing genes with small expressions but high standard errors to be dropped off. Zcut is also able to pick up genes with smaller expressions that have small standard errors, but it does not pick up very small mean expression values, as observed with the log-transform. Thus the new Zcut is more flexible, but there is still a large overlap in detected genes, as we would wish. Of the 3,743 genes identified by Zcut over the weighted data, 2,864 (76.5%) were also identified over the original data. Table 4 records the number of genes detected by each method for the weighted data (α value of .0005).

Interestingly, all of the genes picked by BH are contained in the list of genes picked by Zcut. These are clearly the genes showing large differential expression between B survivors and METS. It is most informative to look at the list of nonoverlapping genes between these two methods. There were 783 genes in this nonoverlappingset. This number was reduced to 779 after genes indicatingpotentialliver or spleen contaminationwere removed. After this, two quality control analyses were run independently (since some samples were prepared in San Francisco and others in Cleveland) to assess differences in samples. Genes showing marked differences in sites were excluded, as were genes showing high variability between tumors on the same patient.

Table 4. Number of Colon Cancer Genes Detected Using the Weighted Regression Approach

Zcut	*FDRmix*	*Bayes exch*	*Z-test*	*Bonf*	*BH*
3,743	3,334	3,433	4,576	1,369	2,960

In the end, there were 476 genes in this nonoverlapping set. Of these, 193 were stripped away from further analysis, because they were expressed sequence tags (ESTs). This left 283 genes, from which EOS probe sets were mapped to Affymetrix U133A and U133B human gene chips. Once this conversion was done, the information was entered into GeneSpring, a program used to extract functional information about genes printed on Affymetrix chips. The remaining genes were then categorized into different functional groups. The findings were quite interesting. Table 5 provides a breakdown of groupings arranged in decreasing biological specificity that represent important functional pathway steps identified in the metastatic process. These include genes that are transcription factors and genes involved in DNA binding, cell adhesion, and various signaling, cell communication, and cascade pathways. In fact, some of the cascade pathways have been identified as potential sources of possible gene–gene interactions due to dimerizations produced during a particular step in a cascade (Cohen et al. 1997).

Table 5. Grouping of Genes Found by Zcut and Not by BH by Biological Function

Functional group	*Number of genes*
Transcription factor	3
DNA binding	1
Cell adhesion	3
G-protein signaling	2
STAT cascade	3
TGF receptor ligand	2
Growth factor receptor ligand	1
Oncogene	2
Intracellular signaling	4
Signal transduction	6
Cell communication	15

8. DISCUSSION

We have shown how to cast the problem of searching for differentially expressed genes in microarrays as a high-dimensional subset selection problem using BAM, a powerful new Bayesian model selection technique. Although we considered the BAM method in detail in the context of the two-way ANOVA model, the method permits extensions to more than two groups. Extensions to ANCOVA models are also possible for use in study designs where gene expressions are measured under various experimental conditions. An important feature of BAM is that it can be used differently depending on the user's needs. For example, Zcut, a model estimator computed from posterior means, amounts to a weighted combination of ridge regression estimators and strikes a balance between low FDR values and higher power. Should the FDR be of greater importance, we have provided a novel technique for estimating the distribution of gene effects under the null hypothesis using a two-component mixture model, leading to the FDRmix procedure. The shrinkage effect of BAM provides the power gains seen for our model estimators. BAM adaptively shrinks test statistics for intermediate- and small-sized effects, whereas larger effects yield test statistics similar to Z-test. This is the reason why even though more gene effects are detected as significant using the new approach than, say, BH, a source of conservatism is built in, minimizing the potential of too many falsely rejected null hypotheses.

The significance of the increased power was evident in our attempt to determine a metastatic genetic signature of colon cancer, known to be a very complex disease process. Zcut detected more than 700 significant genes than BH did. Many of these genes in fact turned out to be potentially implicated in the metastasis of B-survivor solid tumors from the colon to the liver. It is interesting that no currently known genes with obvious involvement in colon cancer metastasis were part of our nonoverlapping list. The implications of this work become more clear; by omitting potentially important genes at the level of initial filtering, further derived discriminant rules based on these filtered subsets of genes may end up leaving out valuable information.

The colon cancer example also illustrates the difficulties in finding global variance-stabilizing transformations for the data. Complex nonlinear relationships between variances and means can result in an adverse affect on the mean when such transformations are applied. This was further illustrated via Poisson simulations. As an alternative, we provided a weighted regression approach. This approach is not limited to the BAM technique and can be used with standard methods as well.

Another issue that can affect inference in these problems are outliers. Ideally, datasets should be trimmed appropriately before analysis, but this is not always an attractive alternative for the practitioner. We have not studied the effects of outliers, because it is beyond the scope of this article. A careful study of their effects is needed. Clearly, some additional robustification of the methods would be required, but this is something that we plan to report on in future work.

APPENDIX A: EXTENSIONS

A.1 More Than Two Groups

Model (4) can be easily extended to the case in which we have more than two groups. For definiteness, we outline the case for three groups. As before, assume that group $l = 1$ is the control group. Groups $l = 2$ and $l = 3$ represent two distinct treatments. If $Y_{j,k,l}$ are the expressions, then testing for a gene-treatment effect can be formulated using the ANOVA model,

$$Y_{j,k,l} = \theta_j + \mu_{j,2} I\{l = 2\} + \mu_{j,3} I\{l = 3\} + \epsilon_{j,k,l},$$
$$j = 1, \dots, p, \quad k = 1, \dots, n_{j,l}, \quad l = 1, 2, 3,$$

where $\epsilon_{j,k,l}$ are iid normal$(0, \sigma^2)$. Testing whether genes differ in groups 2 and 3 from the control corresponds to testing whether $\mu_{j,2}$ and $\mu_{j,3}$ differ from 0. As before, many of the θ_j parameters will be non-zero; thus we reduce the dimension of the parameter space from $3p$ to $2p$ by centering the responses. Applying the appropriate rescaling as part of the preprocessing of the data, we replace $Y_{j,k,l}$ with

$$\widetilde{Y}_{j,k,l} = (Y_{j,k,l} - \overline{Y}_{j,1}) \times \sqrt{n/\hat{\sigma}_n^2},$$

where

$$\hat{\sigma}_n^2 = \frac{1}{n-3p} \sum_{j,k,l} (Y_{j,k,l} - \overline{Y}_{j,1} I\{l = 1\} - \overline{Y}_{j,2} I\{l = 2\}$$
$$- \overline{Y}_{j,3} I\{l = 3\})^2$$

and $\overline{Y}_{j,l}$ is the mean for gene j over group l.

A.2 Differing Measurement Errors: Heteroscedasticity

The ANOVA model (4) can also be extended to handle differing measurement error variances σ_j^2 for $j = 1, \dots, p$. Suppose that genes can be bunched into C clusters based on their variances. Let $i_j \in \{1, \dots, C\}$ indicate the variance cluster gene j belongs to. Now modify (4) by replacing $\epsilon_{j,k,l}$ with variables, say $\tilde{\epsilon}_{j,k,l}$, where $\tilde{\epsilon}_{j,k,l}$ are independent with a normal$(0, \sigma_{i_j}^2)$ distribution. To handle heteroscedastic variances, we apply weighted regression by reweighting observations by the inverse of their standard deviation. Thus (4) can now be written as

$$\sigma_{i_j}^{-1/2} Y_{j,k,l} = \theta_j^* + \mu_j^* I\{l = 2\} + \epsilon_{j,k,l},$$
$$j = 1, \dots, p, \quad k = 1, \dots, n_{j,l}, \quad l = 1, 2,$$

where $\theta_j^* = \sigma_{i_j}^{-1/2} \theta_j$ and $\mu_j^* = \sigma_{i_j}^{-1/2} \mu_j$ denote our new parameters. The C distinct variances $\sigma_1^2, \dots, \sigma_C^2$ are unknown, but they can be estimated accurately when p is large, and thus we can accurately approximate $\sigma_{i_j}^{-1/2} Y_{j,k,l}$. Hence, if $\hat{\sigma}_c^2$ is our estimator for σ_c^2 (we use the usual unbiased estimator), then we simply replace the data $Y_{j,k,l}$ with rescaled data $\hat{\sigma}_{i_j}^{-1/2} Y_{j,k,l}$ and apply the various methods as before.

APPENDIX B: PROOFS

Proof of Theorem 1

With a little bit of rearrangement in (6), we can write

$$(\hat{\theta}_{j,n}, \hat{\mu}_{j,n})^T = M_{j,n} \sqrt{n_{j,2}} (\overline{Y}_{j,2} - \overline{Y}_{j,1}), \qquad \text{where}$$

$$M_{j,n} = \frac{n \widehat{\Sigma}_{j,n}^{-1}}{\hat{\sigma}_n \sigma^2} \Big(\sqrt{n_{j,2}/n_j}, 1 \Big)^T. \tag{B.1}$$

Because the third moment of $\epsilon_{j,k,l}$ is bounded, we can apply the Liapounov central limit theorem (Chow and Teicher 1978, chap. 9.1) to each of the group averages $\overline{Y}_{j,l}$. Use the facts that the averages are independent and that $\sqrt{n_{j,2}/n_{j,1}} \to \sqrt{r_{j,0}}$ to deduce that under the null, $\sqrt{n_{j,2}}(\overline{Y}_{j,2} - \overline{Y}_{j,1}) \overset{\mathrm{d}}{\rightsquigarrow}$ normal$(0, \sigma_0^2(1 + r_{j,0}))$. Meanwhile, a little algebra shows that

$$\hat{\sigma}_n^2 = \frac{1}{n-2p} \sum_{j,k,l} \epsilon_{j,k,l}^2 - \frac{1}{n-2p} \sum_{j,l} n_{j,l} \bar{\varepsilon}_{j,l}^2,$$

where $\bar{\varepsilon}_{j,l} = \sum_{k=1}^{n_{j,l}} \epsilon_{j,k,l}/n_{j,l}$. A bounded fourth moment implies that the first term on the right converges in probability to σ_0^2, while for the second term, a second moment ensures that $\bar{\varepsilon}_{j,l}^2 \xrightarrow{\text{p}} 0$ for each j and l. Hence it follows that $\hat{\sigma}_n^2 \xrightarrow{\text{p}} \sigma_0^2$. Therefore, because $\sigma^2/n \to 1$, it follows that

$$M_{j,n} \xrightarrow{\text{p}} \sigma_0^{-1} \Sigma_j^{-1} \Big(\sqrt{r_{j,0}/(1+r_{j,0})}, 1\Big)^T .$$

Putting the pieces together and appealing to Slutsky's theorem gives the desired result. Note that the degeneracy of the limit poses no problem by applying the Cramér–Wold device.

Proof of Theorem 2

Let $I_0 = \{j : \mu_{j,0} = 0\}$ denote the indices for the 0 coefficients. Define p_0 to be the cardinality of I_0. Because errors are assumed to be normally distributed, (8) implies that $Z_{j,n} \overset{\mathcal{D}}{=} Z_j C_n$, where $C_n = \sigma_0/\hat{\sigma}_n$ and Z_j are independent normal$(\mu_{j,0}/S_{0,n}, 1)$ variables where $S_{0,n} = \sigma_0\sqrt{1/n_{j,1} + 1/n_{j,2}}$. It follows that

$$\begin{aligned} R_F(\alpha) &= \sum_{j\in I_0} \Pr\{|Z_{j,n}| \geq z_{\alpha/2}\} + \sum_{j \in I_0^c} \Pr\{|Z_{j,n}| < z_{\alpha/2}\} \\ &= p_0 \Pr\{|\text{normal}(0,1)| \geq C_n^{-1} z_{\alpha/2}\} + \sum_{j\in I_0^c} \Pr\{|Z_j| < C_n^{-1} z_{\alpha/2}\}. \end{aligned}$$

Recall that $(\hat{\theta}_{j,n}, \hat{\mu}_{j,n})^T$ can be written as (B.1). It can be shown that (remember $\sigma^2 = n$)

$$\hat{\sigma}_n M_{j,n} \to (0, v_{2j}^2/(1+v_{2j}^2))^T, \qquad \text{as } v_{2j-1}^2 \to 0.$$

Because $\hat{\mu}_{j,n}^* = \hat{\mu}_{j,n}\sqrt{n_{j,1}/n_j}$, we use (B.1) and the definition of $Z_{j,n}$ to deduce that for each $0 < \xi_j < 1$, we can find a v_{2j-1}^2 and v_{2j}^2 such that $\hat{\mu}_{j,n}^* = Z_{j,n}\xi_j$. In particular, this means that for each $0 < \delta_1 < 1$ and $0 < \delta_2 < 1$, we can find (v_{2j-1}^2, v_{2j}^2) such that $\hat{\mu}_{j,n}^* = Z_{j,n}\delta_1$ for $j \in I_0$ and $\hat{\mu}_{j,n}^* = Z_{j,n}\delta_2$ for $j \in I_0^c$. Therefore,

$$\begin{aligned} &R_F(\alpha) - R_B(\alpha) \\ &\quad = p_0(\Pr\{|\text{normal}(0,1)| \geq C_n^{-1} z_{\alpha/2}\} \\ &\qquad - \Pr\{|\text{normal}(0,1)| \geq \delta_1^{-1} C_n^{-1} z_{\alpha/2}\}) \\ &\qquad + \sum_{j\in I_0^c} (\Pr\{|Z_j| < C_n^{-1} z_{\alpha/2}\} - \Pr\{|Z_j| < \delta_2^{-1} C_n^{-1} z_{\alpha/2}\}). \end{aligned}$$

Both sums on the right side are continuous functions of α, and each has a minimum and maximum over $\alpha \in [\delta, 1-\delta]$. In particular, the second sum can be made arbitrarily close to 0 uniformly over $\alpha \in [\delta, 1-\delta]$ by choosing δ_2 close to 1, whereas the first sum remains positive and uniformly bounded away from 0 over $\alpha \in [\delta, 1-\delta]$. Thus, for a suitable δ_2, $R_F(\alpha) - R_B(\alpha) > 0$ for each $\alpha \in [\delta, 1-\delta]$.

[Received July 2002. Revised December 2002.]

REFERENCES

Barbieri, M. M., and Berger, J. O. (2002), "Optimal Predictive Model Selection," Discussion Paper 02-02, ISDS.

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.

Brown, P., and Botstein, D. (1999), "Exploring the New World of the Genome With DNA Microarrays," *Nature Genetics*, 21, (suppl), 33–37.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (2000), "Microarray Expression Profiling Identifies Genes With Altered Expression in HDL-Deficient Mice," *Genome Research*, 10, 2022–2029.

Chow, Y. S., and Teicher, H. (1978), *Probability Theory: Independence, Interchangeability, Martingales*, New York: Springer-Verlag.

Cohen, A., Minsky, B., and Schilsky, R. (1997), "Cancer of the Colon," in *Cancer: Principles and Practice of Oncology*, (5th ed.), eds. V. T. J. De-Vita, S. Hellman, and S. Rosenberg, Philadelphia, PA: J. B. Lippincott, pp. 1144–1196.

Crollius, H. R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., Saurin, W., and Weis-Senbach, J. (2000), "Estimate of Human Gene Number Provided by Genome-Wide Analysis Using Tetraodon Nigroviridis DNA Sequence," *Nature Genetics*, 25, 235–238.

Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. M. (2002), "A Variance-Stabilizing Transformation for Gene Expression Microarray Data," *Bioinformatics*, 18, S105–S110.

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

Ewing, B., and Green, P. (2000), "Analysis of Expressed Sequence Tags Indicates 35,000 Human Genes," *Nature Genetics*, 25, 232–234.

Gelman, A., and Tuerlinckx, F. (2000), "Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures," *Computational Statistics*, 15, 373–390.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 85, 398–409.

Genovese, C., and Wasserman, L. (2002a), "Operating Characteristics and Extensions of the FDR Procedure," *Journal of the Royal Statistical Society*, Ser. B, 64, 499–518.

——— (2002b), "False Discovery Rates," Technical Report 762, Carnegie Mellon University, Dept. of Statistics.

Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

Ibrahim, J. G., Chen, M.-H., and Gray, R. J. (2002), "Bayesian Models for Gene Expression With DNA Microarray Data," *Journal of the American Statistical Association*, 97, 88–99.

Ishwaran, H., and Rao, J. S. (2000), "Bayesian Nonparametric MCMC for Large Variable Selection Problems," unpublished manuscript.

Kerr, M., Martin, M., and Churchill, G. (2000), "Analysis of Variance for Gene Expression Microarray Data," *Journal of Computational Biology*, 7, 819–837.

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1036.

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

Rao, J. S., and Bond, M. (2001), "Microarrays: Managing the Data Deluge," *Circulation Research*, 88, 1226–1227.

Rocke, D. M., and Durbin, B. (2001), "A Model for Measurement Error for Gene Expression Arrays," *Journal of Computational Biology*, 8, 567–569.

Schena, M., and Davis, R. W. (1999), "Genes, Genomes and Chips," in *DNA Microarrays: A Practical Approach*, ed. M. Schena, Oxford, U.K.: Oxford University Press.

Schena, M., Heller, R. A., Theriault, T. P., Konrad, K., Lachenmeier, E., and Davis, R. W. (1998), "Microarrays: Biotechnology's Discovery Platform for Functional Genomics," *Trends in Biotechnology*, 16, 301–306.

Speed, T. (2000), "Always Log Spot Intensities and Ratios," Speed Group Microarray Page, at *http://www.stat.berkeley.edu/users/terry/zarray/Html/log.html.*

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Ser. B, 64, 479–498.

Storey, J. D., and Tibshirani, R. (2001), "Estimating False Discovery Rates Under Dependence, With Applications to DNA Microarrays," Technical Report 2001-28, Stanford University, Dept. of Statistics.

Thomas, J. G., Olson, J. M., Tapscott, S. J., and Zhao, L. P. (2001), "An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles," *Genome Research*, 11, 1227–1236.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001), "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proceedings of the National Academy of Sciences of the U.S.A.*, 98, 5116–5121.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001), "Assessing Gene Significance From cDNA Microarray Expression Data via Mixed Models," *Journal of Computational Biology*, 8, 625–637.