Spike and Slab Gene Selection for Multigroup Microarray Data

Hemant ISHWARAN and J. Sunil RAO

DNA microarrays can provide insight into genetic changes that characterize different stages of a disease process. Accurate identification of these changes has significant therapeutic and diagnostic implications. Statistical analysis for multistage (multigroup) data is challenging, however. ANOVA-based extensions of two-sample *Z*-tests, a popular method for detecting differentially expressed genes in two groups, do not work well in multigroup settings. False detection rates are high because of variability of the ordinary least squares estimators and because of regression to the mean induced by correlated parameter estimates. We develop a Bayesian rescaled spike and slab hierarchical model specifically designed for the multigroup gene detection problem. Data preprocessing steps are introduced to deal with unique features of microarray data and to enhance selection performance. We show theoretically that spike and slab models naturally encourage sparse solutions through a process called *selective shrinkage*. This translates into oracle-like gene selection risk performance compared with ordinary least squares estimates. The methodology is illustrated on a large microarray repository of samples from different clinical stages of metastatic colon cancer. Through a functional analysis of selected genes, we show that spike and slab models identify important biological signals while minimizing biologically implausible false detections.

KEY WORDS: Colon cancer; Hypervariance; Penalization; Rescaling; Risk misclassification; Shrinkage; Sparsity; Stochastic variable selection; Zcut.

1. INTRODUCTION

Many invasive diseases, such as cancers, undergo significant transformations during their life spans. Staging of cancers is based on the extent of anatomical invasion from the primary site of development. However, although cancers have well-defined morphological evolution corresponding to clinical stage, very little is known about molecular changes that characterize the stages; this is particularly true for colon cancer; the focus of our present application. High-throughput microarray technology provides a unique opportunity to study this problem. DNA microarrays provide a snapshot of the simultaneous expression of thousands of mRNA transcripts at a given point in time via a single assay. To a first approximation, DNA microarray data provide information about a cell's proteomic composition (using nuclear RNA instead), and thus biological insight into what functional genomic changes might have taken place across the spectrum of a disease. (See Nguyen, Arpat, Wang, and Carroll 2002 for a general overview of biological and technical aspects of microarrays.)

At the same time, microarray data pose a serious statistical challenge due to the sheer volume of information being processed. It is the norm to see data collected on tens of thousands of gene expressions from only a small handful of tissue samples. Data analysis is further complicated because of heterogeneity of variances and correlation of gene expressions due to biological effect or technological artifact. Because it is expected that most genes show no differential gene expression across disease states, the potential for type I errors or false detections is large. For two-group problems, a common strategy is to control the false discovery rate (FDR) using the method of Benjamini and Hochberg (1995) or empirical Bayes methods (Efron, Tibshirani, Storey, and Tusher 2001; Tusher, Tibshirani, and Chu 2001; Storey 2002). However, although these methods work well in controlling FDR, the price paid is often a conservativeness that leads to missing important genes (Ishwaran and Rao 2003). Indeed, in two-group problems, the total number of misclassified genes can be derived in closed form under normality (Genovese and Wasserman 2002, thm. 3). Such calculations suggest that when the fraction of differentially expressing genes is relatively low, misclassification will be high unless FDR is controlled at a high value, thus defeating the purpose of such control.

1.1 Multigroup Data

Multigroup data refers to microarray data collected over different experimental conditions, such as from distinct stages of a disease process. Because of the many questions that could be asked when analyzing such data, most approaches start by simplifying the problem into a composite question that can be tested using a one-dimensional test statistic for each gene. Although this strategy is certainly convenient (e.g., making it possible to apply standard error control methods such as the FDR), it may not be optimal for several reasons. First, the underlying test statistic is likely to be fairly elementary, and thus highly variable, because it will not be *regularized*, that is, constructed in a way that carefully uses information across all genes and samples. Regularization is an important concept in microarray settings, where sample sizes are small. Second, composite statistics are seriously limited in the information that they provide. Consider an analysis using contrasts to check for a specific pattern of differential expression across groups. For example, consider a gene that differentially expresses early on in a disease process such as colon cancer, significantly affecting the biological milieu making it possible for other genes to act, but then later vanishes in terms of biological effect. We call this a *hit-and-run* hypothesis. A contrast, or set of contrasts, looking for hit-and-run genes would simply provide what is equivalent to

Hemant Ishwaran is Associate Staff, Department of Quantitative Health Sciences, 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, two referees, and the associate editor for their diligent review and time spent in helping to improve the manuscript. The authors also thank Sanford Markowitz and Petra Platzer of CWRU for invaluable discussions and assistance in the colon cancer analysis. Hemant Ishwaran was supported by National Science Foundation grant DMS-04-05675. J. Sunil Rao was supported by National Institutes of Health career grant K25-CA89867 and National Science Foundation grant DMS-04-05072.

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214505000000051

a p value for rejecting a null hypothesis of no such pattern being present. What it would not identify is which genes among all such interesting genes are most likely to be truly off for the remainder of the biological process.

Beyond such conventional approaches, very little research seems to have been directed toward the multigroup problem. One notable exception is the recent work by Kendziorski, Newton, Lan, and Gould (2003), which used parametric empirical Bayes method to compute posterior odds for group expression patterns. (There are approximately $g!$ of these if there are g groups.) This requires that gene expression values be exchangeable from a common null and alternative distribution adequately approximated as either a gamma or a lognormal mixture distribution. Genes are classified into a group pattern according to maximum posterior odds. However, although classifying genes into hit-and-run categories is relatively straightforward with a posterior odds approach, it might be difficult to optimally rank genes within a category.

1.2 Contributions and Outline of the Article

Recently, Ishwaran and Rao (2003) introduced a method for detecting differentially expressing genes between two biological groups, termed *Bayesian ANOVA for microarrays* (BAM). This method recasts the statistical problem as a high-dimensional variable selection problem and uses a specific Bayesian hierarchical model geared toward adaptive shrinkage. Using model averaging, a way of accounting for model uncertainty, BAM provides gene effect estimates shrunken relative to standard least squares estimates in which primarily only the nondifferentially expressing gene effects are shrunken. This a general phenomenon that we call *selective shrinkage*, which plays a crucial role in our extension of BAM to multigroup data.

This extension differs in some subtle but important ways from the original methodology. One key innovation is our use of orthogonality. We show how to cast the multigroup microarray problem in terms of an ANOVA framework and then transform the problem into a high-dimensional orthogonal model after a simple dimension-reduction and rescaling step. The transformed data are then modeled using a Bayesian rescaled spike and slab model as introduced by Ishwaran and Rao (2005). Besides leading to computational simplifications, orthogonality is a key ingredient in establishing selective shrinkage and other theoretical properties of the method. These results, outlined shortly, provide a deeper understanding of the methodology than those of Ishwaran and Rao (2003). Another important advancement is our ability to systematically deal with heterogeneity of variances across genes. We show how a weighted regression clustering technique used in tandem with graphical diagnostic plots can effectively deal with this problem without resorting to global transformations that can distort signal to noise ratios.

Sections 3–5 contain the theoretical underpinnings of our methodology. Section 6 illustrates the methodology on a large database involving colon cancer. The article concludes with a discussion in Section 7. Our key contributions and primary findings are summarized as follows:

1. Approaching the problem through an ANOVA framework is advantageous, because it allows estimation of all gene-group differential effects simultaneously and avoids having to resort to pairwise group comparisons or user-constructed contrasts to identify interesting gene expression profiles across experimental groups.

2. False detection rates for least squares based test statistics are inflated due to a regression to the mean effect in multigroup problems. This effect, which is due to correlation between test statistics, is mitagated using a rescaled spike and slab approach because of the effects of shrinkage.

3. Under a suitably chosen bimodal hypervariance prior, we are able to achieve a selective shrinkage property in which Bayesian test statistics are large when the differential effect is real (Corollary 1) and shrunk to zero for differential effects that are zero, with this latter effect quantified using an exact representation (Thm. 3). These Bayesian test statistics can be viewed as solutions to a least squares constrained optimization problem in which each gene-group differential parameter has a unique penalty term that is adaptively estimated from the data.

4. We demonstrate that selective shrinkage is sufficient for oracle-like uniformly low-risk misclassification (Thm. 2) and that risk performance improves with sparsity of the parameter space. Given that many gene-group differential parameters are expected to be zero, this suggests that misclassification performance of rescaled spike and slab models naturally improves in multigroup problems.

5. We derive an adaptive data-driven graphical cutoff rule. We characterize the rule by showing that it has the property of optimally (asymptotically) separating genes that are differentially expressing from those that are not (Thm. 4). This shows that the method is risk consistent.

6. We show how rescaled spike and slab models can be used to pull out complex patterns of differential gene expression across stages of colon cancer and liver metastasis that include the patterns described earlier as well as many other interesting patterns. A detailed biological functional analysis of selected genes provides insight into pathways that are activated or deactivated across stages of this disease. Graphical plots for simultaneously visualizing all stagewise differential effects are introduced.

2. RESCALED SPIKE AND SLAB MODELS FOR MULTIGROUP DATA

The data used in illustrating our approach come from a large microarray repository of colon cancer samples of various stages collected at the Ireland Cancer Center of Case Western Reserve University. All gene expression data were compiled using high-density 59K-on-one gene chips developed by EOS Biotechnology. These are Affymetrix-derived chips with proprietary probe sets. The high density of probe sets reflects known genes and ESTs (expressed sequence tags), as well as predicted exons.

Consider Figure 1, which is part of a detailed analysis given later in Section 6. The figure is based on data from four distinct colon tissue samples: Duke's B, C, and D and liver METS, as expressed by the Astler–Coller–Duke staging system (Cohen, Minsky, and Schilsky 1997). The Duke B's in our dataset were actually Duke BSurvivors, comprising patients still alive

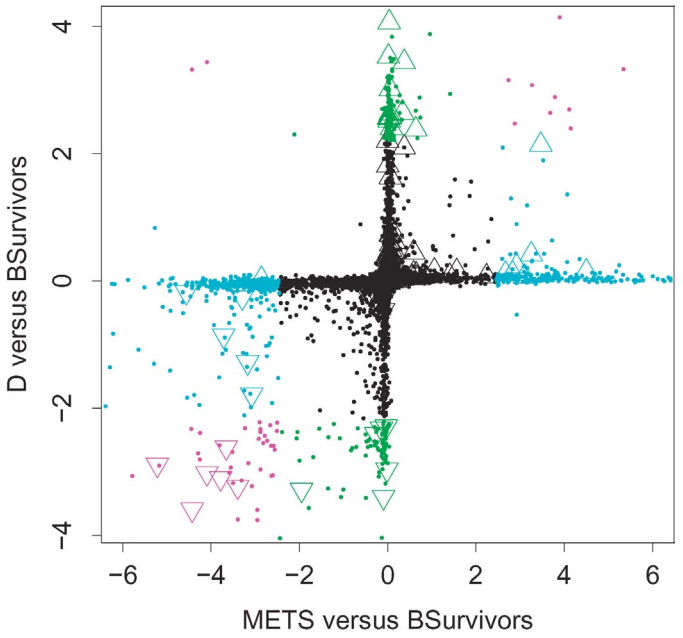

Figure 1. Zcut Values From Colon Cancer Analysis. (Some values are not shown because plot is zoomed in.) Vertical and horizontal axes are tests for difference between D's versus BSurvivors and METS versus BSurvivors. Genes differentially expressing for both groups (magenta); D's but not METS (green); METS but not D's (blue); neither group (black). Also indicated are C versus BSurvivors differentially expressed genes by △ (turning on) and ▽ (turning off).

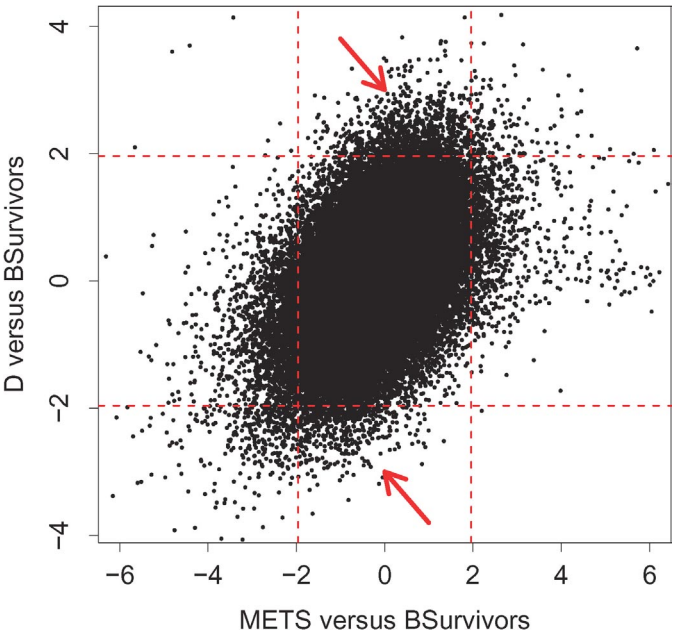

Figure 2. Z-Test Statistics Using a Traditional ANOVA Model. Arrows indicate quadrants containing potential hit-and-run genes using 95% confidence regions.

from the time of initial diagnosis and represent an intermediate stage of cancer. Duke's C tumors represent a progressive worsening of the disease as the cancer begins to spread deeper into the colon wall from the innermost tissue layers, and also to nearby lymph nodes. Liver METS (METS) represent the most advanced stage of the disease where the tumor has metastasized to a distant site, in this case the liver (the other major site is the lung). Duke's D tumors correspond to the tumor depost remaining in the primary organ site after metastasis.

Plotted in Figure 1 are Bayesian estimated differential gene effects (defined later as *Zcut values*) for comparing the METS and D's with the BSurvivors (x- and y-axes). Also overlayed on the plot are triangles for identifying genes turning off or on for stage C relative to the BSurvivors. In the figure we have used color to highlight stagewise gene effects of biological interest. Points colored in magenta are genes with significant differential expression across the D's and METS being either turned on or turned off relative to the BSurvivors. For example, the small cluster of magenta triangles in the bottom-left quadrant indicate genes that turn off throughout the C, D, and METS samples. Data points colored in green and blue indicate genes that are significant (in either direction) for only the stage D's and not the METS or for only the METS and not the stage D's. In particular, green points that hug the y-axis are those exhibiting significant changes from BSurvivors to D's but whose METS expression resemble the BSurvivors. These are hit-and-run genes, mentioned in Section 1, and are of particular importance, because they have a very specific early effect only.

Least squares estimates (Z-tests) from a standard ANOVA model provides a strikingly different plot (Fig. 2). Especially apparent is the ellipsoid nature of the figure. As we show later (Sec. 3), this is due to a regression to the mean effect caused by the correlation between the Z-statistics for the METS versus BSurvivors and the D's versus BSurvivors. Regression to the mean inflates false detections and makes it difficult to delineate signal from noise. Notice how difficult it is to identify any hit-and-run candidates. For example, early hit-and-run genes might be the ones in the quadrants indicated by the two arrows, but this is not so clear.

2.1 Multigroup ANOVA Models

A common recurring theme in this article is the distinction between the underlying model assumed for the data and the Bayesian hierarchical model used for inference. In the former case we assume a nonparametric ANOVA framework, whereas in the latter case we use what we call a rescaled spike and slab hierarchical model. Figure 1, and the detailed analysis of Section 6, were based on this model. Rescaled spike and slab models, discussed in detail in Section 2.4, involve a normal hierarchy, which at first glance might seem quite at odds with the nonparametric framework assumed for the data. However, as we show in Section 3, the rescaled spike and slab hierarchy induces a gene differential parameter estimate that can be viewed as a penalized least squares solution, and thus is a fully nonparametric estimate. Moreover, we show that these estimates have certain properties making them quite useful in multigroup microarray settings under weak assumptions to the data. Hence, although the rescaled spike and slab models make use of a normal hierarchy, which might appear to make an implicit distributional statement about the data, in fact the resulting estimators have properties that hold under minimal distributional assumptions.

We begin by describing the ANOVA model for the data. Let gene expression values be denoted by $Y_{i,j}$, where $i = 1, \ldots, n$ and $j = 1, \ldots, M$. The value $Y_{i,j}$ represents the expression value for gene j from the ith microarray chip. To keep track

of group membership, each chip i is assigned a group label $\mathscr{G}_i \in \{1, \ldots, g\}$.

To search for genes exhibiting a differential effect, we define a baseline group against which change in expression levels is measured. It is convenient notationally to let g be the baseline group. In the colon cancer analysis, BSurvivors represent the baseline group $g = 4$, whereas stage C, D, and liver METS tissues are group labels 1, 2, and 3. The BSurvivors represent a natural baseline because our interest is in studying how colon cancer progresses from this intermediate stage. However, a different baseline group could certainly be used if we were interested in a different biological question.

Let $\mathbb{I}(\cdot)$ denote the indicator function, and let $n_k = \#\{i : \mathscr{G}_i = k\}$ denote the number of samples from group k. The total sample size for each gene is $n = \sum_{k=1}^{g} n_k$. The multigroup ANOVA model is defined as

$$Y_{i,j} = \underbrace{\theta_j}_{\substack{\text{baseline effect}\\ \text{for gene } j}} + \underbrace{\sum_{k=1}^{g-1} \beta_{k,j} \mathbb{I}\{\mathscr{G}_i = k\}}_{\substack{\text{gene-group interaction}\\ \text{(differential effect)}}} + \underbrace{\varepsilon_{i,j}}_{\substack{\text{independent error,}\\ \mathscr{D}_{i,j}(0,\sigma_j^2)}}, \quad i = 1, \ldots, n,\ j = 1, \ldots, M. \quad (1)$$

Note that importantly, we take a distribution-free approach to the data. We assume only that $\varepsilon_{i,j}$ are independent such that $\mathbb{E}(\varepsilon_{i,j}) = 0$ and $\mathbb{E}(\varepsilon_{i,j}^2) = \sigma_j^2$.

The multigroup ANOVA model does not assume an equal variance model (i.e., $\sigma_j^2 = \sigma_0^2$), because this will be unrealistic for microarray data. Often microarray data exhibit a complex relationship between the mean and the standard deviation, with standard deviations usually increasing with means. In contrast, although an equal variance model is unrealistic, a model that has one variance σ_j^2 for each gene will lack power. What we need is a way to group variances into clusters with each cluster having a unique value, but using as few clusters as possible. This is an important example of regularization. Section 2.2 introduces a weighted regression clustering approach for just this kind of regularization.

In (1), each gene has a parameter θ_j representing a mean effect for the baseline group g. The value for this parameter is of little scientific interest. Section 2.3 uses a dimension-reduction step to remove its effect from the model. Of interest are the parameters $\beta_{1,j}, \ldots, \beta_{g-1,j}$, which measure the difference in mean expression value relative to the baseline. These parameters represent a gene-group interaction effect and are used to test for differential expression. A non-0 value for $\beta_{k,j}$ indicates a relative change in mean expression value. A positive value for $\beta_{k,j}$ indicates an increase in the value (gene j is "turning on" for group k relative to baseline); a negative value indicates a relative decrease in the value (the gene is "turning off").

We apply a sequence of three preprocessing steps that convert (1) into a format suitable for a Bayesian hierarchical analysis:

(P1) A weighted regression technique is applied to transform the data so as to satisfy an equal variance assumption $\sigma_j^2 = 1$. This technique avoids problems associated with global variance-stabilizing transformations such as log-transformations (see Durbin, Hardin, Hawkins, and Rocke 2002; Ishwaran and Rao 2003 for problems with using log-transformations). Furthermore, it does not change the signal-to-noise ratio of the data for a gene (see Sec. 2.2).

(P2) The dimension of the model is reduced (Sec. 2.3). Observe that (1) has gM parameters. However, as indicated, because our interest is only in identifying non-0 $\beta_{k,j}$ parameters, the significant gene-group effects, it is convenient to force θ_j to be 0. To do so, we replace $Y_{i,j}$ by the centered value $Y_{i,j}^+ - \overline{Y}_{g,j}^+$, where $\overline{Y}_{g,j}^+ = \sum_{\{i:\mathscr{G}_i = g\}} Y_{i,j}^+ / n_g$ and the superscript "+" is used to indicate data transformed under (P1). The analysis is then restricted to observations with group labels $\mathscr{G}_i \neq g$. This reduces the dimension of the problem from Mg to $M(g-1)$. After centering $Y_{i,j}$, we also rescale the observations by a factor equal to the square root of the overall sample size. This rescaling has the effect of acting like a penalty term and is needed for optimal risk performance.

(P3) The transformed data is modeled using an orthogonal rescaled spike and slab model (Sec. 2.4).

2.2 Weighted Regression Transformation (Step P1)

To transform the data, we cluster genes according to their pooled standard deviations, $\widehat{\sigma}_j$, defined by

$$\widehat{\sigma}_j^2 = \frac{1}{n-g} \sum_{k=1}^{g} \sum_{\{i:\mathscr{G}_i = k\}} (Y_{i,j} - \overline{Y}_{k,j})^2,$$

where $\overline{Y}_{k,j} = \sum_{\{i:\mathscr{G}_i = k\}} Y_{i,j}/n_k$. Genes are clustered according to preset percentile values for $\widehat{\sigma}_j$. For example, to create $\mathscr{C} = 2$ clusters, genes are clustered by whether $\widehat{\sigma}_j$ is less than or equal to the 99th percentile, and for $\mathscr{C} = 3$ clusters, genes are clustered according to the 95th and 99th percentiles. This can be repeated for any number of clusters $\mathscr{C} = 1, \ldots, M$. The extreme case where $\mathscr{C} = 1$ corresponds to the original data (i.e., untransformed), whereas $\mathscr{C} = M$ corresponds to a unique cluster for each gene.

After the $\mathscr{C}$ clusters are identified, each observation in a cluster is scaled by dividing by the square root of the mean sample variance. If $\mathscr{J}_l$ are the indices for genes in cluster l, then the mean sample variance for cluster l is

$$\mathscr{M}_l = \frac{1}{M_l} \sum_{j \in \mathscr{J}_l} \widehat{\sigma}_j^2 = \frac{1}{(n-g)M_l} \sum_{j \in \mathscr{J}_l} \sum_{k=1}^{g} \sum_{\{i:\mathscr{G}_i = k\}} (Y_{i,j} - \overline{Y}_{k,j})^2,$$

where M_l is the number of genes in the cluster. Because observations in a cluster are multiplied by the same value, all expression values for a gene j are multiplied by the same value, independent of their group membership. Therefore, this type of transformation has the important property that it does not affect the signal-to-noise ratio for a gene.

Remark 1. Used in a linear regression model, this type of transformation can also be viewed as a weighted regression technique. In this case $\sqrt{\mathscr{M}_l}$ act as the weights in the model (see Ishwaran and Rao 2003 for further discussion).

The scaling-transformation is designed so that for the transformed data, $\widehat{\sigma}_j^2 = 1$ for each j when $\mathscr{C} = M$. But this will "overfit" the data. Even if an equal variance model is true (i.e., $\sigma_j^2 = 1$), then we still expect some variability in $\widehat{\sigma}_j^2$ around the value of 1. Therefore, rather than choosing a large value of $\mathscr{C}$ and potentially overregularizing the problem with a subsequent loss in power, the preferable method is to increase $\mathscr{C}$ gradually until an equal variance model is satisfied.

We introduce two graphical techniques to accomplish this. The first technique is a percentile standard deviation graph; the second is what we call a *V-plot*. Figures 3 and 5 illustrate their application to our data. Figure 3 graphs the 1st–99th percentile values for $\widehat{\sigma}_j$ following a particular transformation. Transformations corresponding to clusters of size $\mathscr{C} = 2$ to $\mathscr{C} = 100$ were considered. ($\mathscr{C} = 1$, the untransformed data, was excluded because its range of values were so different it could not be overlayed on the figure.) When $\mathscr{C} = 2, \ldots, 5$, the data are in gross violation of an equal variance model. However, things quickly improve after this. With only $\mathscr{C} = 8$ clusters (formed by clustering the 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles), we find that most genes have pooled standard deviations near 1. There is also a high overlap between the observed percentiles and those derived under a normal equal variance model [computed assuming $\varepsilon_{i,j}$ are iid N(0, 1)]. After adding only a few more clusters, we see a marked departure between the observed percentiles and the equal variance benchmark. When $\mathscr{C}$ increases, the pooled standard deviations become uniformly closer to 1, but we lose power as regularization decreases. The best choice for the number of clusters appears to be $\mathscr{C} = 8$.

Remark 2. Our strategy for choosing the percentiles used for clustering was to work systematically from higher to lower values. Doing so allows us to get the most regularization with the least number of clusters. However, the rule used for choosing the specific percentile value is admittedly somewhat arbitrary and it is useful to have a more objective approach. One rule that could be used is based on a binary splitting approach defined as follows. For $\mathscr{C} = 2$, group genes by whether their standard deviations are less than or greater than the Pth percentile. Compute the variance of the standard deviations in both groups. Take their weighted average (weighted by sample size) and divide by the overall variance of the standard deviation for all genes. This is a measure of within to between variability. Find the value for P, call it P^*, that minimizes the within to between measure of variability. This is the percentile used for $\mathscr{C} = 2$. For $\mathscr{C} = 3$, repeat the same procedure on all genes whose standard deviations are in the $[0, P^*]$ percentile group. Let P^{**} be the value that minimizes the within to between variability in this subset. The percentile groups for $\mathscr{C} = 3$ are $[0, P^{**}]$, $(P^{**}, P^*]$, and $(P^*, 1]$. For $\mathscr{C} = 4$, repeat the process on the $[0, P^{**}]$ group, and so on. Figure 4 shows how this method compares with our previous technique. In Figure 4(a), one can see that the percentiles for the binary splitting rule are slightly larger. In Figure 4(b), however, we see that this discrepancy makes no difference in the final values for the transformed standard deviations.

The V-plot depicted in Figure 5 should be used in tandem with Figure 3. The V-plot is a graph of the transformed group mean difference $(\overline{Y}_{k,j}^+ - \overline{Y}_{g,j}^+)$ versus the absolute value of the t-test for gene j and group k,

$$t_{k,j} = \frac{\overline{Y}_{k,j}^+ - \overline{Y}_{g,j}^+}{\sqrt{\widehat{\sigma}_{k,j}^2/n_k + \widehat{\sigma}_{g,j}^2/n_g}},$$

where

$$\widehat{\sigma}_{k,j}^2 = \frac{1}{n_k - 1} \sum_{\{i:\mathscr{G}_i = k\}} (Y_{i,j}^+ - \overline{Y}_{k,j}^+)^2.$$

If an equal variance model across groups is true, then these two values should be linearly related with a theoretical slope of $(1/n_k + 1/n_g)^{-1/2}$. The dashed line in the figure depicts what this linear relationship should be. As can be seen, there appears to be good agreement, although some points fall off the theoretical line. Therefore, as an additional check, we compared the

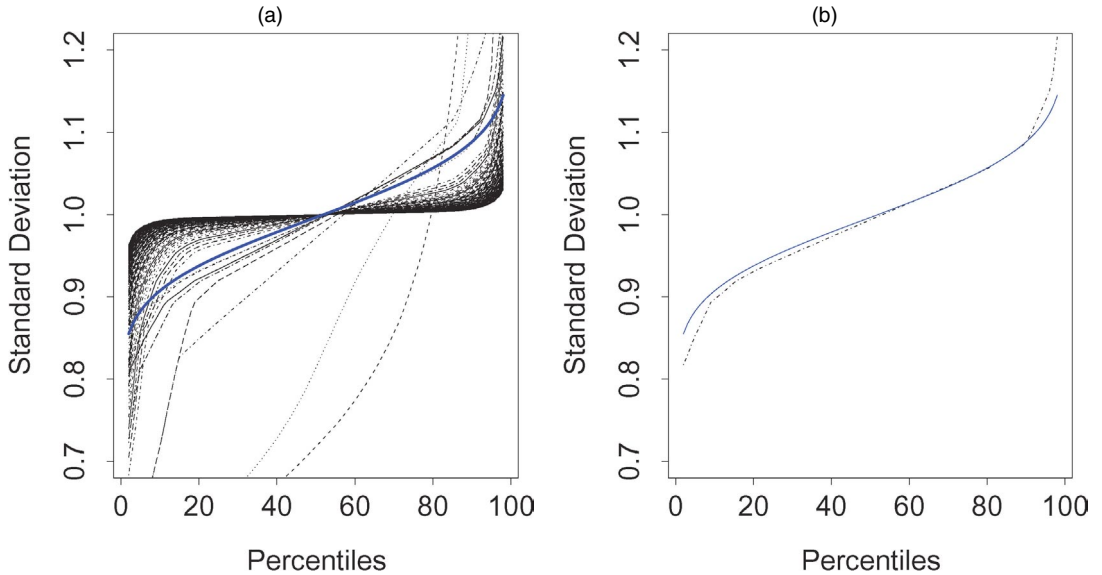

Figure 3. Pooled Standard Deviations for Genes Sorted by Percentile. (a) Transformed data from weighted regression with $\mathscr{C} = 2, \ldots, 100$ clusters ($\mathscr{C} = 2$ is the most vertical curve, whereas $\mathscr{C} = 100$ is nearly horizontal). (b) When $\mathscr{C} = 8$, pooled standard deviations are near 1 and are closely approximated by quantiles derived from normal equal variance model $\sigma_0^2 = 1$ (thick blue line).

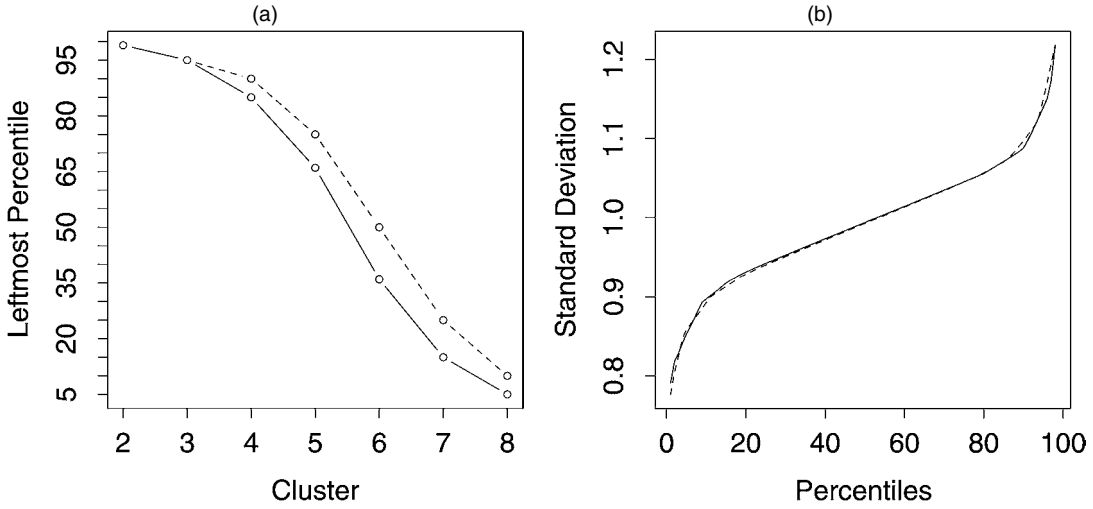

Figure 4. (a) Leftmost Percentile Value as Number of Clusters Increases and (b) Quantiles for Standard Deviations From Transformed Data ($\mathscr{C} = 8$). In both our method (——) and the binary splitting method (– – ––) a new cluster is formed by introducing a new "leftmost" percentile value with which to split the data in terms of the pooled standard deviation.

quantiles for the standard error,

$$\sqrt{\widehat{\sigma}^2_{k,j}/n_k + \widehat{\sigma}^2_{g,j}/n_g}, \tag{2}$$

obtained from the transformed data with those obtained under an iid $\mathrm{N}(0, 1)$ model. Figure 6 shows that these values are in close agreement, further confirming the choice of $\mathscr{C} = 8$ as a satisfactory transformation. Therefore, this was the value that we used throughout.

2.3 Rescaling and Centering the Data (Step P2)

To reduce the dimension of the problem, we center the transformed data from step P1 by the baseline mean value (group g)

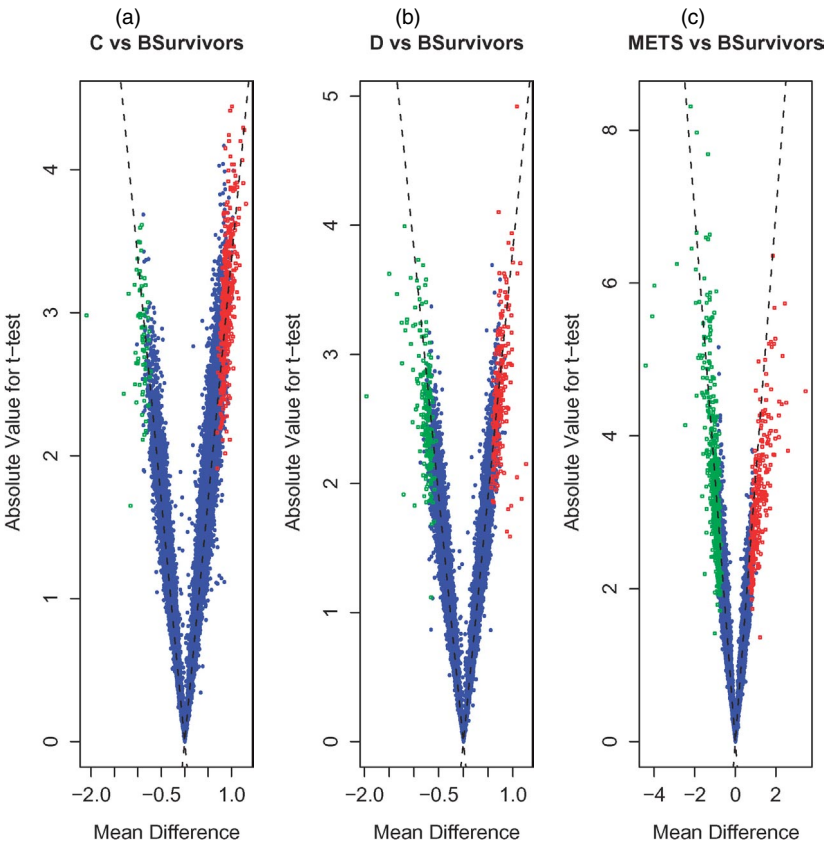

Figure 5. V-Plot. Absolute value for t-test versus group mean difference (transformed data with $\mathscr{C} = 8$): (a) C versus BSurvivors, (b) D versus BSurvivors, (c) METS versus BSurvivors. The dashed line represents the theoretical value if equal variance model $\sigma_0^2 = 1$ holds. Differentially expressing genes are indicated by green (off) and red (on). Nonsignificant genes are blue.

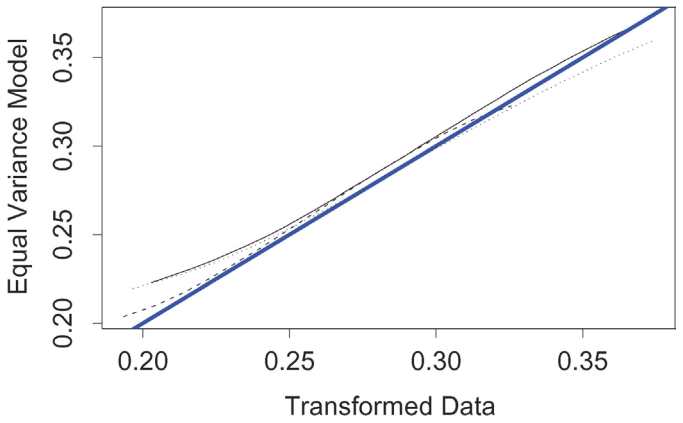

Figure 6. Quantiles for Standard Error (2) for Transformed Data versus Those Obtained Under an Equal Variance Model Across Groups. The solid, medium-dashed, and thin-dashed black lines are C's, D's, and METS. The solid blue line is y = x.

and restrict analysis to groups $k \neq g$. A rescaling is also used to allow the use of a rescaled spike and slab model. (We discuss the reasons for using this type of rescaling shortly; see Sec. 2.5.) In place of $Y_{i,j}^{+}$, we use

$$Y_{i,j}^{*} = \sqrt{\frac{N}{\widehat{\sigma}_N^2}}(Y_{i,j}^{+} - \overline{Y}_{g,j}^{+}), \qquad \mathscr{G}_i \neq g, \tag{3}$$

where $N = (n - n_g)M$ is the total sample size and $\widehat{\sigma}_N^2$ is the usual unbiased estimator for σ_0^2 using all of the data,

$$\widehat{\sigma}_N^2 = \frac{1}{(n-g)M} \sum_{j=1}^{M} \sum_{k=1}^{g} \sum_{\{i:\mathscr{G}_i=k\}} (Y_{i,j}^{+} - \overline{Y}_{k,j}^{+})^2.$$

Any estimator $\widehat{\sigma}_N^2$ could be used as long as it is consistent for σ_0^2 under an assumed equal variance model.

2.4 Rescaled Spike and Slab Models (Step P3)

Following steps P1 and P2, we convert the multigroup ANOVA model to a rescaled spike and slab model as discussed by Ishwaran and Rao (2005). Let $\mathbf{Y}_j^{*} = (Y_{1,j}^{*}, \ldots, Y_{n-n_g,j}^{*})^t$ be the transformed values (3) for gene j. For convenience, we assume that the first n_1 values of $\mathbf{Y}_j^{*}$ have group label 1, the next n_2 observations have group label 2, and so forth. Let $\boldsymbol{\beta}_j = (\beta_{1,j}, \ldots, \beta_{g-1,j})^t$ denote the regression parameters for gene j. The rescaled spike and slab multigroup model is

$$\begin{aligned} (\mathbf{Y}_j^{*} | \boldsymbol{\beta}_j, \sigma^2) &\sim \mathrm{N}(\mathbf{X}_j\boldsymbol{\beta}_j, N\sigma^2\mathbf{I}), \qquad j = 1, \ldots, M, \\ (\boldsymbol{\beta}_j | \boldsymbol{\gamma}) &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}_j), \\ \boldsymbol{\gamma} &\sim \pi(d\boldsymbol{\gamma}), \\ \sigma^2 &\sim \mu(d\sigma^2), \end{aligned} \tag{4}$$

where $\boldsymbol{\Gamma}_j$ is the diagonal matrix with diagonal entries obtained from $\boldsymbol{\gamma}_j = (\gamma_{1,j}, \ldots, \gamma_{g-1,j})^t$. Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^t, \ldots, \boldsymbol{\gamma}_M^t)^t$ be the $M(g-1)$-dimensional hypervariance vector.

The design matrix $\mathbf{X}_j$ for gene j, of dimension $(n - n_g) \times (g - 1)$, is chosen to satisfy orthogonality. Let $m_0 = 0$ and $m_k = \sum_{l=1}^{k} n_l$ for $k = 1, \ldots, g-1$. Then coordinates $m_0 + 1$ through m_1 correspond to observations with group labels equal to 1, coordinates $m_1 + 1$ through m_2 correspond to group labels equal to 2, and so forth. This means that $\mathbf{X}_j = [\mathbf{x}_1, \ldots, \mathbf{x}_{g-1}]$, where $\mathbf{x}_k$ is the vector with 0 everywhere except at coordinates $m_{k-1} + 1$ through m_k, where its values are $(N/n_k)^{1/2}$,

$$\mathbf{x}_k = (N/n_k)^{1/2}(0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)^t.$$

Notice that each $\mathbf{X}_j$ is an orthogonal design matrix, $\mathbf{X}_j^t\mathbf{X}_j = N\mathbf{I}$.

2.5 Penalization

The rescaling (3) of the response by a factor of $\sqrt{N}$ is suggested by work of Ishwaran and Rao (2005). They showed that in the general spike and slab model, rescaling the data controls the amount of penalization that the posterior mean exhibits in relation to the ordinary least squares estimate, and that to achieve a nonvanishing relative effect, the rescaling should be equal to the total sample size. Furthermore, in orthogonal models, setting the rescaling at this level encourages selective shrinkage, which we expect should naturally induce adaptive sparsity in the high-dimensional setting of microarray data. In the context of our model, the effect of rescaling is captured by the term N appearing in the first level of the hierarchy (4). In Section 3.1 we explicitly show how rescaling impacts penalization of the posterior gene differential parameters.

2.6 Zero Prior Means

Notice that the prior for $\boldsymbol{\beta}_j$ in (4) is centered at $\mathbf{0}$. This fits in with our belief that the underlying data are sparse. That is, a priori, we believe that differential gene effects are zero.

2.7 The $Y_{i,j}^{\star}$ Values Are Correlated

Because of the centering used in (3), the measurements $Y_{i,j}^{*}$ are correlated across a gene j. Because of this, it may seem unusual that in (4) we use an independent variance structure for the response. This issue goes back to our point made earlier in Section 2.1, that there is a distinction between the underlying model assumed for the data and the Bayesian hierarchy used for inference. We do not assume that the $Y_{i,j}^{*}$'s are uncorrelated; in fact, correlation plays a key role in model selection performance. We show later (Sec. 3.4) that correlation introduces a regression to the mean effect that inflates the number of false-positive findings using a frequentist Z-test analysis, but that this regression to the mean effect is mitigated because of the effects of shrinkage when using a rescaled spike and slab model.

2.8 The Role of σ^2

Note that because N takes on the role of a penalty effect, the value for σ^2 in (4) assumes the role of an adaptive penalty adjustment. It plays no role in terms of the variance, however. This is because we have rescaled the data by $\widehat{\sigma}_N$, which removes the effect of σ_0^2. Our experience has shown that the posterior for σ^2 is usually concentrated near 1, although some adaptiveness can occur.

2.9 Continuous Bimodal Priors

For $\boldsymbol{\gamma}$, we use the continuous bimodal priors of Ishwaran and Rao (2003, 2005). The prior π for $\boldsymbol{\gamma}$ is induced by the

following parameterization. Define $\gamma_{k,j}$ by $\gamma_{k,j} = I_{k,j}\tau^2_{k,j}$, where $I_{k,j}$ and $\tau^2_{k,j}$ are parameters with priors specified according to

$$\begin{aligned}(I_{k,j}|v_0, w_k) &\overset{\text{iid}}{\sim} (1-w_k)\delta_{v_0}(\cdot) + w_k\delta_1(\cdot),\\ (\tau^{-2}_{k,j}|a_1, a_2) &\overset{\text{iid}}{\sim} \text{gamma}(a_1, a_2),\\ w_k &\overset{\text{iid}}{\sim} \text{uniform}[0,1],\\ k = 1,\ldots,g-1,\ j &= 1,\ldots,M.\end{aligned} \tag{5}$$

The choice for v_0 (a small near 0 value) and a_1 and a_2 (the shape and scale parameters for a gamma density) are selected so that $\gamma_{k,j}$ has a continuous bimodal distribution with a spike at v_0 and a right-continuous tail (Fig. 7). Such a prior allows the posterior to either shrink a coefficient toward 0, or not.

The parameters w_k also play a special role. Because w_k controls the probability that $I_{k,j}$ equals 1, it behaves as a *complexity parameter*, controlling the number of genes found to be differentially expressed for group k. We use $g-1$ complexity parameters, one parameter for each group. It is possible to use a shared complexity w for all groups, but our experience has shown this to be less robust in multigroup problems. For example, if the expression values for a nonbaseline group is significantly different than measurements from other nonbaseline group measurements, then this can unduly inflate or shrink the overall value for w. Using a unique complexity for each group is also important when group membership has a natural ordering, as in our data setting. For colon cancer, group membership corresponds to stage of disease, and gene activity is expected to (generally) increase as the disease worsens. Therefore, we would expect to see some kind of increasing pattern for complexity parameter values as group membership, k, increases. Our analysis of Figure 1 used a unique complexity parameter for each of the groups: stage C, stage D, and liver METS. We found the posterior means for w_1, w_2, and w_3 to be .008, .008, and .0132. This indicates a significant increase in the overall number of genes found to be significant for the METS but there is roughly similar behavior in terms of total gene involvement for the C's and D's. We discuss this point again in Section 6.

Remark 3. Throughout the article, we use the hyperparameter values of Ishwaran and Rao (2003, 2005) of $v_0 = .005$, $a_1 = 5$, and $a_2 = 50$. (Fig. 7 is the density under these choices.) For σ^2, we used an inverse gamma prior,

$$\sigma^{-2} \sim \text{gamma}(b_1, b_2),$$

where $b_1 = b_2 = .0001$. All parameters in the rescaled spike and slab model are estimated in a Gibbs sampling scheme, called *stochastic variable selection* (SVS) (see Ishwaran and Rao 2005 for details).

Remark 4. The use of a continuous bimodal prior is one of the unique features that distinguishes our spike and slab method from other popular spike and slab approaches, such as those discussed by George and McCulloch (1993). (See Ishwaran and Rao 2005 for a more thorough discussion on these differences.)

3. THE EFFECT OF SHRINKAGE IN MULTIGROUP MICROARRAY DATA

In this section we explicitly work out the posterior mean from our rescaled spike and slab model to understand the effects of shrinkage. We show that the posterior mean from such models can be interpreted as a Bayesian test statistic that is shrunk relative to a frequentist test. The amount of shrinkage is shown to be related to an adaptively estimated penalty term. We then use this Bayesian test statistic in conjunction with a thresholding rule to define a rule for finding differentially expressing genes (Sec. 3.3). Our study of shrinkage also helps to explain its role in reducing regression to the mean (Sec. 3.4). Note that our discussion, and all subsequent work hereafter, tacitly assumes that the underlying data satisfy an equal variance model in which $\sigma^2_j = \sigma^2_0$. Such an assumption should hold with reasonable accuracy on applying the variance-stabilizing transformation of Section 2.2. Henceforth, theory is developed assuming that such a transformation has been made to ensure that (1) holds for the transformed data under an equal variance assumption $\sigma^2_j = \sigma^2_0$. Thus, hereafter we drop the use of our superscript "+" notation.

3.1 A Bayes Test Statistic

The posterior mean is easily derived by using conjugacy and the orthogonality of $\mathbf{X}_j$. One can show that

$$(\boldsymbol{\beta}_j|\boldsymbol{\Gamma}_j, \sigma^2, \mathbf{Y}^*_j) \sim \text{N}(\boldsymbol{\mu}_j, \sigma^2\boldsymbol{\Sigma}_j), \tag{6}$$

where $\boldsymbol{\Sigma}_j = \text{diag}\{v_{1,j}, \ldots, v_{g-1,j}\}$,

$$v_{k,j} = \frac{\gamma_{k,j}}{\gamma_{k,j} + \sigma^2}, \qquad k = 1,\ldots,g-1,$$

and

$$\boldsymbol{\mu}_j = \widehat{\sigma}_N^{-1}\big(v_{1,j}\sqrt{n_1}(\overline{Y}_{1,j} - \overline{Y}_{g,j}), \ldots, v_{g-1,j}\sqrt{n_{g-1}}(\overline{Y}_{g-1,j} - \overline{Y}_{g,j})\big)^t.$$

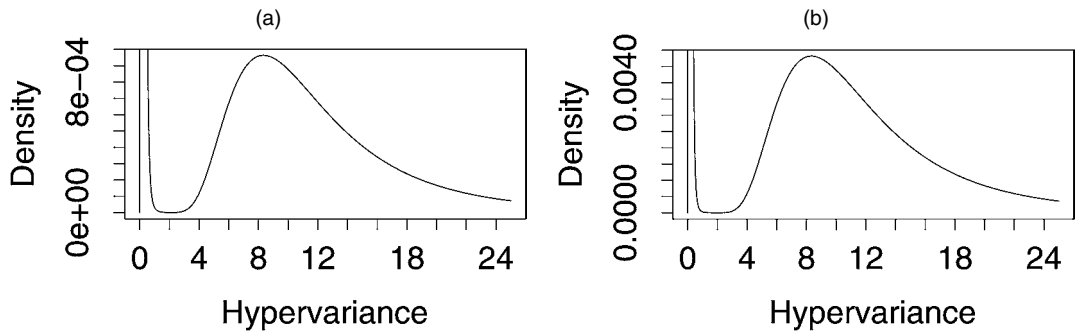

Figure 7. Conditional Density for $\gamma_{k,j}$ Given w_k: (a) $w_k = .01$ and (b) $w_k = .05$. Observe that only the height of the density changes as w_k is varied.

The dimension-reduction step of subtracting $\overline{Y}_{g,j}$ from gene expression values slightly alters the interpretation of the posterior mean as a test statistic. A simple sample size correction is needed to adjust things. Define

$$\widehat{\beta}^*_{k,j} = \left(1 + \frac{n_k}{n_g}\right)^{-1/2} \mathbb{E}(\beta_{k,j}|\mathbf{Y}^*), \qquad k = 1, \ldots, g-1;$$

then, letting $V_{k,j} = \mathbb{E}(\nu_{k,j}|\mathbf{Y}^*)$,

$$\widehat{\beta}^*_{k,j} = V_{k,j} \frac{\overline{Y}_{k,j} - \overline{Y}_{g,j}}{\widehat{\sigma}_N \sqrt{1/n_k + 1/n_g}}, \qquad k = 1, \ldots, g-1,$$

which, up to the factor $V_{k,j}$, is the Z-test statistic from an ANOVA model for comparing the mean for group k to the mean for group g. The value $0 \le V_{k,j} \le 1$ represents a shrinkage factor. Thus $\widehat{\beta}^*_{k,j}$ can be interpreted as a *Bayesian shrinkage test statistic*.

One can also view $\widehat{\beta}^*_{k,j}$ as a solution to a constrained least squares optimization problem in which $V_{k,j}$ are related to the penalties. Let $\widehat{\boldsymbol{\beta}}^*_j = (\widehat{\beta}^*_{1,j}, \ldots, \widehat{\beta}^*_{g-1,j})^t$. It can be shown that

$$\widehat{\boldsymbol{\beta}}^*_j = \mathbf{D}_N \times \arg\min_{\boldsymbol{\beta}_j} \left\{ \frac{1}{N} \|\mathbf{Y}^*_j - \mathbf{X}_j \boldsymbol{\beta}_j\|^2 + \sum_{k=1}^{g-1} \frac{1 - V_{k,j}}{V_{k,j}} \beta^2_{k,j} \right\}, \tag{7}$$

where $\mathbf{D}_N$ is the $(g-1) \times (g-1)$ diagonal matrix with entries $\{(1 + n_k/n_g)^{-1/2} : k = 1, \ldots, g-1\}$. The value for N appearing in (7) is due to the rescaling (3). Note, importantly, how this specific choice ensures that the second term in the expression, the penalization effect due to $V_{k,j}$, does not vanish asymptotically.

Observe how each $\beta_{k,j}$ coefficient in (7) is penalized by a unique value $(1 - V_{k,j})/V_{k,j}$. The closer $V_{k,j}$ is to 1, the smaller the penalty, whereas the closer $V_{k,j}$ is to 0, the larger the penalty. It is clear that the more adaptive $V_{k,j}$ is to the true parameter value, the more accurate the Bayes test statistic will be in finding differentially expressing genes. Given that we expect relatively few gene-group differential effects, an optimal solution to (7) would naturally be sparse; that is, we would expect many parameters to be zero and thus many coefficients to have large penalty terms. Section 4 discusses how sparsity is related to improved risk misclassification. In Section 5, rescaled spike and slab models are shown to be capable of adaptive penalization.

3.2 Limiting Distributions

We also refer to the posterior mean test statistics $\widehat{\beta}^*_{k,j}$ as *Zcut values*. The following theorem states the limiting distribution for the conditional posterior mean. This can be used to calibrate a thresholding rule for Zcut for identifying differentially expressing genes.

Theorem 1. Assume that (1) represents the true model where $\varepsilon_{i,j}$ are independent such that $\mathbb{E}(\varepsilon_{i,j}) = 0$, $\mathbb{E}(\varepsilon^2_{i,j}) = \sigma^2_0$, and $\mathbb{E}(\varepsilon^4_{i,j}) \le A_0$ for some fixed constant $A_0 < \infty$. Assume that $n_1, \ldots, n_g \to \infty$ such that $n_k/n \to \Pi_{k,0} > 0$ for $k = 1, \ldots, g$. Let $\boldsymbol{\beta}^*_j = (\beta^*_{1,j}, \ldots, \beta^*_{g-1,j})^t$, where

$$\beta^*_{k,j} = \nu_{k,j} \frac{\overline{Y}_{k,j} - \overline{Y}_{g,j}}{\widehat{\sigma}_N \sqrt{1/n_k + 1/n_g}}, \qquad k = 1, \ldots, g-1.$$

Then $\boldsymbol{\beta}^*_j \overset{\mathrm{d}}{\rightsquigarrow} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_j \boldsymbol{\Omega} \boldsymbol{\Sigma}^t_j)$ under the null hypothesis $\beta_{1,j} = \cdots = \beta_{g-1,j} = 0$, where

$$\boldsymbol{\Omega} = \begin{pmatrix} 1 & b_{1,0}b_{2,0} & b_{1,0}b_{3,0} & \cdots & b_{1,0}b_{g-1,0} \\ b_{1,0}b_{2,0} & 1 & b_{2,0}b_{3,0} & \cdots & b_{2,0}b_{g-1,0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{1,0}b_{g-1,0} & b_{2,0}b_{g-1,0} & \cdots & \cdots & 1 \end{pmatrix}$$

and $b_{k,0} = (1 + \Pi_{g,0}/\Pi_{k,0})^{-1/2}$.

3.3 The Zcut Rule

Theorem 1 shows that $\beta^*_{k,j} \overset{\mathrm{d}}{\approx} \mathrm{N}(0, \nu^2_{k,j})$ under the null. Because $\nu^2_{k,j}$ is likely to be small when the null is true (Sec. 5 explains why this is true), this shows that comparing Zcut to a $\mathrm{N}(0, 1)$ will result in few rejections when $\beta_{k,j}$ is really 0. In contrast, if $\beta_{k,j}$ is truly non-0, then $\gamma_{k,j}$ is expected to be large and $\nu_{k,j}$ is nearly 1 (again see Sec. 5). Thus Zcut will be nearly equal to a two-sample Z-test with a $\mathrm{N}(0, 1)$ distribution. These arguments suggest that Zcut can be compared with a standard normal distribution to assess significance. We call this the *multigroup Zcut rule*, a generalization of the idea presented by Ishwaran and Rao (2003).

The Multigroup Zcut Rule. Classify gene j as differentially expressing if $|\widehat{\beta}^*_{k,j}| \ge \mathrm{Cut}$. The cutoff value Cut can be set to $z_{\alpha/2}$, the $100 \times (1 - \alpha/2)$ percentile of a standard normal distribution, for some suitably chosen α.

In practice, setting a good α value for Zcut (or any other test statistic) can be difficult given the large number of tests. Alternatively, we show how Cut can be chosen using a data-driven graphical rule that is risk consistent; see Section 5.1.

3.4 Shrinkage Discourages Regression to the Mean

Shrinkage is often thought of as a point estimation phenomenon, but with the help of Theorem 1 we show that shrinkage can also play a key role in reducing the covariance between point estimates. This leads to better performance in terms of misclassification and false discoveries. Another way to view this phenomenon is that without the benefits of shrinkage, one ends up with a regression to the mean effect.

Figure 8, a plot of Z-test statistics for the colon cancer data, illustrates how regression to the mean can inflate false discoveries. The Z-test for comparing group k to the baseline for gene j is defined as

$$\widehat{\beta}_{k,j} = \frac{\overline{Y}_{k,j} - \overline{Y}_{g,j}}{\widehat{\sigma}_N \sqrt{1/n_k + 1/n_g}}. \tag{8}$$

This is, of course, similar to the Zcut value $\widehat{\beta}^*_{k,j}$, but without the benefit of a shrinkage factor. Figure 8 is based on the same data as shown in Figure 2, but now we have indicated genes that exhibit what we call a *sawtooth pattern*. A sawtooth pattern corresponds to a C versus BSurvivors effect, no D versus BSurvivors effect, and an METS versus BSurvivors effect (all significant effects could be in either direction). There currently is no biologically known mechanism that would support such patterns of expression, so we must conclude that these genes represent false positives.

Using a confidence region derived from individual 95% confidence intervals, we count 245 genes with a sawtooth pattern.

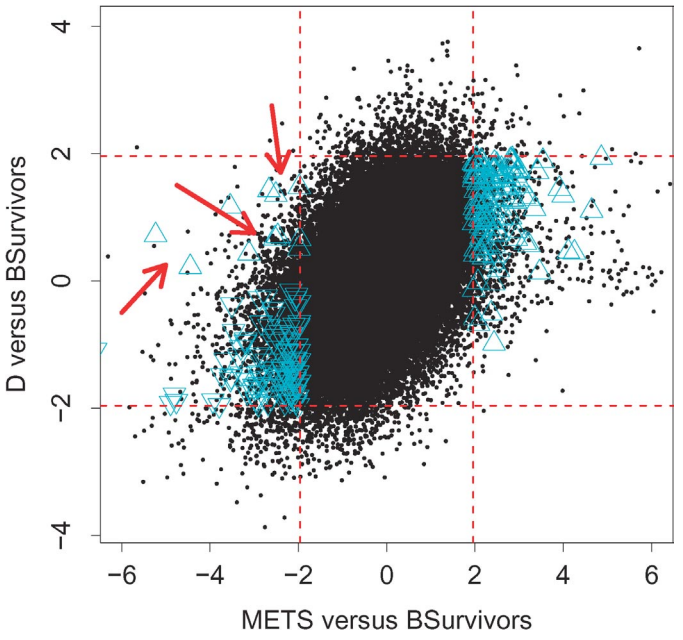

Figure 8. Z-Test Statistics for Colon Cancer Analysis. Triangles indicate sawtooth genes. These are genes that turn on (△) or off (▽) for stage C, do not differentially express for stage D, and then turn back on or off for the METS.

We show that regression to the mean offers a reasonable explanation for this surprisingly large number of false positives. Let groups 1, 2, and 3 refer to the C's, D's, and METS. Then the set of sawtooth genes in Figure 8 are those with indices j in the set

$$\mathscr{R} = \big\{j : |\widehat{\beta}_{1,j}| \geq 1.96, |\widehat{\beta}_{2,j}| < 1.96, |\widehat{\beta}_{3,j}| \geq 1.96, \\ j = 1, \ldots, M\big\}.$$

Each gene in $\mathscr{R}$ represents a misclassification. In total, there are seven different ways that a misclassification can occur depending on the true value for the betas. The most likely type of misclassification are genes with $\beta_{1,j,0} = \beta_{2,j,0} = \beta_{3,j,0} = 0$, where $\beta_{k,j,0}$ denotes the true beta value. These are genes with no differential effect across any group. These should make up the bulk of false positives simply because they represent the vast majority of genes overall and so, by sheer volume, are more likely to be misclassified. We examine this case in detail. For convenience, we call this the $(0, 0, 0)$ case.

To study what can happen to a $(0, 0, 0)$ gene, consider the asymptotic behavior of the Z-tests (8) for the C's, D's, and METS under the null. Under the conditions of Theorem 1, $(\widehat{\beta}_{1,j}, \widehat{\beta}_{2,j}, \widehat{\beta}_{3,j})^t \overset{\mathrm{d}}{\rightsquigarrow} \mathrm{N}(\mathbf{0}, \mathbf{\Omega}_{3\times 3})$, where $\mathbf{\Omega}_{3\times 3}$ is the 3×3 submatrix of $\mathbf{\Omega}$ for the first three coordinates of $\boldsymbol{\beta}$. This suggests that, asymptotically,

$$\mathbb{E}(\widehat{\beta}_{1,j}|\widehat{\beta}_{2,j}, \widehat{\beta}_{3,j}) = (\rho_{1,2}, \rho_{1,3}) \begin{pmatrix} 1 & \rho_{2,3} \\ \rho_{2,3} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\beta}_{2,j} \\ \widehat{\beta}_{3,j} \end{pmatrix} \\ = \frac{\rho_{1,2} - \rho_{1,3}\rho_{2,3}}{1 - \rho_{2,3}^2}\widehat{\beta}_{2,j} + \frac{\rho_{1,3} - \rho_{1,2}\rho_{2,3}}{1 - \rho_{2,3}^2}\widehat{\beta}_{3,j},$$

where $0 < \rho_{k,l} = b_{k,0}b_{l,0} < 1$. Therefore, if $\widehat{\beta}_{3,j}$ is large enough so that $|\widehat{\beta}_{3,j}| \geq 1.96$ (which constitutes a type I error, an event that may occur frequently when M is large), then even if $\widehat{\beta}_{2,j}$ is nonsignificant, $\widehat{\beta}_{1,j}$ can still be large enough so that $|\widehat{\beta}_{1,j}| \geq 1.96$. This is an example of regression to the mean. In this case a relatively large value for $\widehat{\beta}_{1,j}$ is induced by a large value for $\widehat{\beta}_{3,j}$ due to the positive correlation between the two estimates. Thus for the sawtooth genes, the C's may be significant simply due to a spurious significant effect for the METS.

Remark 5. This argument explains the sawtooth pattern seen for the genes in Figure 8 highlighted by triangles except for the small cluster indicated by thick arrows. Notice that for these genes, the direction of differential expression for the C's is in the *opposite* direction of the METS. These are still false positives, and they are still likely to be an artifact of regression to the mean, but the explanation is slightly different. Most likely, the positive C effect is just noise, and because of the correlation with the D's, the effect due to the D's is also positive, but is regressed to the mean to a sufficient degree to make it nonsignificant.

In comparison, consider what happens under the $(0, 0, 0)$ case with Zcut. By the Cauchy–Schwartz inequality,

$$\operatorname{cov}(\widehat{\beta}^*_{1,j}, \widehat{\beta}^*_{3,j}) \leq \sqrt{\operatorname{var}(\widehat{\beta}^*_{1,j})\operatorname{var}(\widehat{\beta}^*_{3,j})}.$$

Observe that

$$\operatorname{var}(\widehat{\beta}^*_{1,j}) \leq \mathbb{E}(\widehat{\beta}^{*2}_{1,j}) = \mathbb{E}\big\{\mathbb{E}(\beta^{*2}_{1,j}|\mathbf{Y}^*)\big\} \\ \lessapprox \sqrt{\mathbb{E}(\mathrm{N}(0,1)^4)\mathbb{E}(V^4_{1,j})} \\ \leq \sqrt{\mathbb{E}(\mathrm{N}(0,1)^4)\mathbb{E}(V_{1,j})},$$

where the approximation on the right side is suggested by Theorem 1 and the Cauchy–Schwartz inequality. Later theory (Sec. 5) will show that $\mathbb{E}(V_{1,j})$ is small when $\beta_{1,j,0} = 0$, and consequently (using a similar argument to bound the variance of $\widehat{\beta}^*_{3,j}$),

$$\operatorname{cov}(\widehat{\beta}^*_{1,j}, \widehat{\beta}^*_{3,j}) \approx 0.$$

This shows that the Zcut values $\widehat{\beta}^*_{1,j}$ and $\widehat{\beta}^*_{3,j}$ should be approximately asymptotically independent, thus greatly diminishing the possibility of a regression to the mean effect. Moreover, because $\widehat{\beta}^*_{1,j}$, $\widehat{\beta}^*_{2,j}$, and $\widehat{\beta}^*_{3,j}$ will be small (by virtue of their variances being small), we would expect far fewer $(0, 0, 0)$ genes in $\mathscr{R}$ relative to the frequentist Z-test method. In fact, in comparison, Figure 1 contained only 13 genes with a sawtooth pattern.

4. ORACLE RISK PERFORMANCE

The following result shows the extent to which shrinkage can enhance identification of differentially expressing genes. We consider a simplified rescaled spike and slab multigroup model in which the hypervariance $\mathbf{\Gamma}_j$ is fixed at some value. We show that there exists an oracle value for $\mathbf{\Gamma}_j$ leading to superior risk performance compared with the analogous frequentist estimator, and that this performance is naturally improved in the presence of sparsity.

For the following argument, we assume a simplified version of (4),

$$(\mathbf{Y}^*_j|\boldsymbol{\beta}_j) \sim \mathrm{N}(\mathbf{X}_j\boldsymbol{\beta}_j, N\mathbf{I}), \qquad j = 1, \ldots, M, \\ (\boldsymbol{\beta}_j|\boldsymbol{\gamma}_j) \sim \mathrm{N}(\mathbf{0}, \mathbf{\Gamma}_j). \tag{9}$$

Notice that in (9) we have removed the prior for $\boldsymbol{\gamma}$ and fix σ^2 at a value of 1. To measure performance in detecting differentially

expressing genes, we introduce the following measure of risk. Let $\delta_{k,j} \in \{0, 1\}$ be the binary value recording whether gene j is truly differentially expressed over group k with respect to the baseline group g. Thus $\delta_{k,j} = 0$ if $\beta_{k,j,0} = 0$; otherwise, $\delta_{k,j} = 1$ if $\beta_{k,j,0} \neq 0$, where $\beta_{k,j,0}$ is the true value for $\beta_{k,j}$. Let $\widehat{\delta}^*_{k,j}(C)$ be our decision rule based on Zcut for some cutoff value C; that is, $\widehat{\delta}^*_{k,j}(C) = 1$ if and only if $|\widehat{\beta}^*_{k,j}| \geq C$. The total number of misclassifications for gene j is

$$\text{total misclassifications}_j(C) = \sum_{k=1}^{g-1} \mathbb{I}\{\widehat{\delta}^*_{k,j}(C) \neq \delta_{k,j}\}. \tag{10}$$

Call the expected value of (10) the risk for Zcut for gene j.

The frequentist analog of Zcut is the Z-test statistic $\widehat{\beta}_{k,j}$ defined by (8). To study how well Zcut stacks up against Z-test, let $\widehat{\delta}_{k,j}(C)$ be the decision rule based on Z-test; that is, $\widehat{\delta}_{k,j}(C) = 1$ if and only if $|\widehat{\beta}_{k,j}| \geq C$. Define the risk for Z-test similar to (10). The following theorem, a generalization of a theorem of Ishwaran and Rao (2003), shows that there exists a hypervariance such that Zcut has *uniformly* better risk performance than Z-test.

Theorem 2. Assume that (1) holds where $\varepsilon_{i,j}$ are iid $\mathrm{N}(0, \sigma_0^2)$. If not all $g-1$ groups are differentially expressing for gene j, then for each C_1 and C_2 such that $0 < C_1 < C_2 < \infty$, there exists a $\mathbf{\Gamma}_j$ for (9) such that

$$\sum_{k=1}^{g-1} \mathbb{P}\{\widehat{\delta}^*_{k,j}(C) \neq \delta_{k,j}\} < \sum_{k=1}^{g-1} \mathbb{P}\{\widehat{\delta}_{k,j}(C) \neq \delta_{k,j}\}$$

for any C, where $C_1 \leq C \leq C_2$.

Theorem 2 applies only to genes for which at least one group has no differential effect. This should not be of concern in practice, though. Typically, only a fraction of genes will exhibit differential expression, and of those, only a small fraction will be differentially expressing over all groups. That is, we expect the underlying model to be sparse. However, one should not interpret a requirement of sparsity to mean that Zcut would be outperformed by Z-test otherwise. For genes differentially expressing over all groups, Zcut's risk will be almost identical to Z-test's. This is a consequence of the selective shrinkage capability of the posterior, a topic that we discuss in the next section. So Zcut does not suffer in nonsparse settings; however, in sparse settings, risk performance is amplified.

The overall performance of a procedure is measured by its *total risk*, defined as the risk over all genes (Ishwaran and Rao 2003). Using a similar proof as for Theorem 2, we can show under the same conditions there exists a $\mathbf{\Gamma} = (\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_M)$ such that for almost all C,

$$\sum_{j=1}^{M} \sum_{k=1}^{g-1} \mathbb{P}\{\widehat{\delta}^*_{k,j}(C) \neq \delta_{k,j}\} < \sum_{j=1}^{M} \sum_{k=1}^{g-1} \mathbb{P}\{\widehat{\delta}_{k,j}(C) \neq \delta_{k,j}\}, \tag{11}$$

as long as there is at least one gene that is not differentially expressing over all its groups. Note by our previous discussion that the inequality (11) becomes more pronounced when many genes are not differentially expressing over all groups. This makes Zcut very attractive in sparse multigroup settings.

5. SELECTIVE SHRINKAGE

A careful inspection of the proof of Theorem 2 shows that uniform total risk performance relies on the oracle $\mathbf{\Gamma}$ being selected so that its values are small for zero $\beta_{k,j}$ coefficients and large for nonzero coefficients. Because the value for the hypervariance controls the amount of shrinkage of Zcut, this shows that the oracle $\mathbf{\Gamma}$ is chosen precisely to induce selective shrinkage of Zcut. Thus selective shrinkage is a sufficient condition for optimal risk performance.

Theorem 2 relies on knowledge of the oracle value, which is unknown in practice, and it requires normality. We show that selective shrinkage can be achieved without such assumptions under our rescaled spike and slab framework. We begin by presenting a closed-form expression for the conditional posterior mean of $\nu_{k,j}$ (what can be thought of as a standardized hypervariance). Note that although this result is stated for spike and slab priors using group-specific complexity parameters, it also holds under shared complexity models. Note also that we assume $\sigma^2 = 1$, to simplify some technical arguments in the proof. But this is a reasonable assumption, because the posterior for σ^2 is naturally concentrated around 1.

Theorem 3. Consider a rescaled spike and slab model (4) where $\sigma^2 = 1$ with a continuous bimodal prior (5) with hyperparameters set as in Remark 3. Let $\nu_{k,j} = \gamma_{k,j}/(\gamma_{k,j} + 1)$. If $\mathbb{E}^*(\cdot|w_k)$ denotes the conditional posterior expectation given w_k, then

$$\mathbb{E}^*(\nu_{k,j}|w_k) = \frac{\int_0^1 \nu \exp(\nu\xi_{k,j}^2)(1-\nu)^{-3/2} f_{k,j}(\nu/(1-\nu)|w_k)\, d\nu}{\int_0^1 \exp(\nu\xi_{k,j}^2)(1-\nu)^{-3/2} f_{k,j}(\nu/(1-\nu)|w_k)\, d\nu}, \tag{12}$$

where $f_{k,j}(\cdot|w_k) = (1-w_k)g_0(\cdot) + w_k g_1(\cdot)$ is the prior density for $\gamma_{k,j}$ given w_k, $g_0(u) = v_0 u^{-2} g(v_0 u^{-1})$, $g_1(u) = u^{-2} g(u^{-1})$,

$$g(u) = \frac{a_2^{a_1}}{(a_1-1)!} u^{a_1-1} \exp(-a_2 u),$$

and $\xi_{k,j} = 2^{-1/2} \widehat{\sigma}_N^{-1} \sqrt{n_k}(\overline{Y}_{k,j} - \overline{Y}_{g,j})$.

Theorem 3 provides considerable insight into the behavior of Zcut and shows that this estimator has a selective shrinkage property that holds in a nonparametric sense in finite samples. This is a direct consequence of (7). Recall that this relationship explicity characterizes Zcut in terms of the penalty term $\mathbb{E}^*(1-\nu_{k,j})/\mathbb{E}^*(\nu_{k,j})$, where $\mathbb{E}^*$ denotes expectation under the posterior. The representation of Zcut as a solution to an L_2 optimization problem shows that it can be viewed as a nonparametric estimator. The exact amounts of penalization and how this translates into selective shrinkage for finite samples is quantified by the theorem.

Consider the expression $\xi_{k,j}^2$ appearing in (12). This term represents the gene-differential effect for group k (relative to the baseline) and directly controls how much the prior is influenced by the data. In particular, unless $\xi_{k,j}^2$ is unduly large, the posterior mean value for $\nu_{k,j}$ should be relatively close to the value under the prior, which in turn should be reasonably small if w_k is small. Thus $\mathbb{E}^*(\nu_{k,j})$ will be small and Zcut penalized and shrunk toward zero. But if the differential effect $\xi_{k,j}^2$ is large, then $\mathbb{E}^*(\nu_{k,j})$ is nearly 1 and there is hardly any penalization,

and Zcut will hardly be shrunk at all relative to the frequentist estimate. This indicates a selective shrinkage property for Zcut. From our discussion of Theorem 2, we know that this can have profound implications for total risk performance, especially in sparse settings when heavy shrinkage occurs over the many zero coefficients.

5.1 Shrinkage Plots and Risk Consistency

Theorem 3 specifies not only the finite-sample behavior of $\mathbb{E}^*(\nu_{k,j}|w_k)$, but also its asymptotic behavior as the sample size increases, at least in the case when the true parameter $\beta_{k,j,0}$ is 0. (This is a direct consequence of the continuous mapping theorem.) Figure 9(a) demonstrates this effect. This figure graphs the posterior density for the standardized hypervariance $\nu_{k,j}$ against values of $\xi^2_{k,j}$ spanning the various percentiles of a χ^2_1 distribution. This is the limiting distribution of $\xi^2_{k,j}$ under the null $\beta_{k,j,0} = 0$ in a balanced design in which group sizes are equal (i.e., $\Pi_{k,0}/\Pi_{g,0} = 1$). All plots are based on a value of $w_k = .01$, which is approximately the value estimated for all

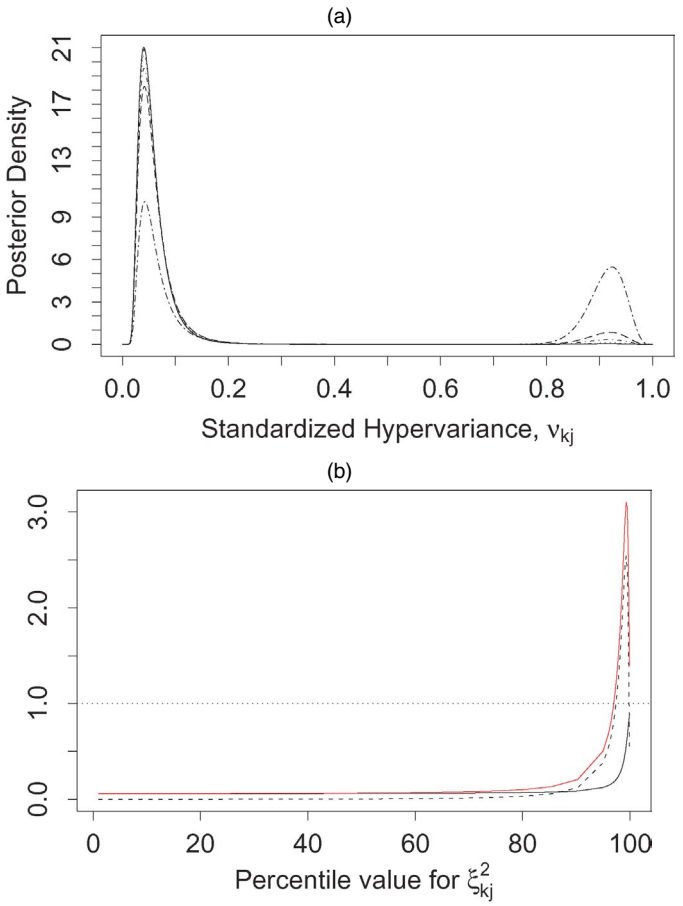

Figure 9. Hypervariance Properties. (a) Posterior density for $\nu_{k,j}$ conditioned on $w_k = .01$ for various values of $\xi^2_{k,j}$. Values for $\xi^2_{k,j}$ are selected from the 25th, 50th, 75th, 90th, 95th, and 99th percentiles of its asymptotic distribution under the null for a balanced design. The density mode on the right increases as $\xi^2_{k,j}$ increases, whereas the mode on the left increases as $\xi^2_{k,j}$ decreases. (b) The solid black line is $\mathbb{E}^*(\nu_{k,j})$, the dashed black line is $2\xi^2_{k,j}$ var$^*(\nu_{k,j})$, and the solid red line is var$^*(\beta_{k,j})$. All values are estimated assuming that $w_k = .01$ is fixed. The horizontal dashed line at 1 is the theoretical limit of $\mathbb{E}^*(\nu_{k,j})$ and var$^*(\beta_{k,j})$ as $\xi^2_{k,j} \to \infty$.

three groups in our colon cancer analysis. The figure shows that the posterior density is bimodal, with the relative heights of the modes controlled by the value of $\xi^2_{k,j}$. However, unless $\xi^2_{k,j}$ is extremely large, the posterior mean for $\nu_{k,j}$ will be relatively small. Notice how this figure hints at the significant gains in total risk performance that might be occurring in our data setting. The amount of shrinkage is very pronounced due to the small value of the complexity w_k.

To consider the asymptotic behavior when $\beta_{k,j,0} \neq 0$, we introduce the following corollary. We use this result as a first step in establishing the risk consistency of Zcut.

Corollary 1. Assume the same conditions as in Theorems 1 and 3. If $\beta_{k,j,0} \neq 0$, then $\mathbb{E}^*(\nu_{k,j}) = 1 + O_p(n_k^{-1})$.

We establish risk consistency for Zcut under a data-adaptive graphical method called a *shrinkage plot*, a graphical device introduced by Ishwaran and Rao (2003) for data-adaptively selecting cutoff values for Zcut. A shrinkage plot is a plot of the Zcut value $\widehat{\beta}^*_{k,j}$ against the posterior variance for $\beta_{k,j}$. Differentially expressing genes are identified as those with posterior variances coalescing near the value of 1 and with large Zcut values.

We begin by using Corollary 1 to give a heuristic explanation for why a shrinkage plot works. A rigorous proof, and risk consistency for Zcut, follow afterward. Let var$^*(\cdot)$ denote the variance under the posterior. From (6), assuming that $\sigma^2 = 1$,

$$\text{var}^*(\beta_{k,j}) = \mathbb{E}^*(\nu_{k,j}) + 2\xi^2_{k,j}\,\text{var}^*(\nu_{k,j}). \tag{13}$$

By Corollary 1, we know that $\mathbb{E}^*(\nu_{k,j}) \xrightarrow{\text{p}} 1$ for truly nonzero coefficients, and thus the behavior of var$^*(\beta_{k,j})$ around the value of 1 will depend on the second term in (13). This second term, $\xi^2_{k,j}\,\text{var}^*(\nu_{k,j})$, reflects a trade-off between signal from the data and the amount of shrinkage induced by the posterior. The effect from the data, $\xi^2_{k,j}$, converges to infinity when $\beta_{k,j} \neq 0$, but at the same time var$^*(\nu_{k,j})$ is expected to converge to 0. If these two terms converge at a rate such that $\xi^2_{k,j}\,\text{var}^*(\nu_{k,j}) \to 0$, then var$^*(\beta_{k,j}) \to 1$. This suggests that to find genes most likely to be truly differentially expressed, we should choose those genes with large Zcut values and whose posterior variances coalesce near 1. This is the underlying premise for the shrinkage plot. Figure 9(b) provides an illustration of the interplay between the terms in (13).

We now formalize this argument and show that Zcut using this adaptive cutoff is risk-consistent. This result also implicitly establishes a rate of convergence for the posterior variance of $\nu_{k,j}$.

Theorem 4. Under the conditions of Corollary 1,

$$\left(|\widehat{\beta}^*_{k,j}|,\, \text{var}^*(\beta_{k,j})\right) \xrightarrow{\text{p}} (\infty, 1) \quad \text{as } n \to \infty$$

if and only if $\beta_{k,j,0} \neq 0$. Furthermore, var$^*(\nu_{k,j}) = O_p(n_k^{-2})$ if $\beta_{k,j,0} \neq 0$.

Theorem 4 shows that the nonzero coefficients will form a cloud of points on the shrinkage plot that in the limit drifts off to $\pm\infty$ on the left and right sides while converging to the value of 1 along the y-axis. Choosing genes in this cloud of points will identify *all* nonzero coefficients, and thus will ensure that the risk is zero in the limit. This shows that Zcut is risk-consistent under this selection mechanism.

6. MULTIGROUP SIGNATURES OF COLON CANCER

We now return to a more complete analysis of the colon cancer data discussed earlier. As of July 2003 there were 104 samples in our database, of which 25 were BSurvivors, 21 were Duke C, 35 were Duke D, and 23 were liver METS. [Note that these samples are not part of the data used by Ishwaran and Rao (2003), but rather a fresh set.] Using earlier notation, the group sizes were $n_1 = 21$, $n_2 = 35$, $n_3 = 23$, and $n_4 = 25$ for each of the $M = 59{,}618$ probe sets. Figure 10 presents shrinkage plots showing gene effects estimated by Zcut from the rescaled spike and slab model versus posterior variances. The three plots represent individual stagewise comparisons to the BSurvivors that represent the baseline stage. The red-colored points indicate genes that have been significantly turned on, and the green-colored points indicate those that have been significantly turned off. Blue coloring indicates no difference in gene expression relative to the BSurvivors. A gene was determined to be significant using an eyeball technique of setting the Zcut cutoff value to coincide with posterior variances coalescing at a value of 1. There were 350, 338, and 774 genes significant for the C, D, and METS groups. The slightly larger number of genes found to be significant for the C's when compared to the D's, and then the large jump in the number of significant genes for the METS, can be explained biologically. To be classified as stage C, the tumor must have entered the middle layers of the colon wall and started to innervate the neighboring lymphatic system. In some sense this represents a milieu of the "tumor being revved up." A Duke's D has more lymphatic involvement, but many of the genes involved in the initial revving-up process may "shut down." Given the distant metastasis site of the liver METS from the colon, there is an implication that another revving-up process may take place.

A more interesting microstructure is attained by mapping each gene to a particular differential expression pattern across the groups—for instance, a gene that was turned on across all stages or one that might have been turned off for a subset of the intermediate stages and show no effect otherwise (i.e., a hit-and-run gene). In this particular dataset there would be 27 such patterns, some of which would clearly be biologically implausible. Table 1 summarizes how genes are mapped to mutually exclusive pattern types. With this dataset, we found 12 plausible patterns of interest. A gene-specific pattern is represented by a vector (d_1, d_2, d_3), where $d_k \in \{0, 1, -1\}$ depending on if group k's gene expression level was unchanged, up-regulated (gene turned on), or down-regulated (gene turned off) relative to the BSurvivors.

Probesets listed in Table 1 were mapped to the GO database to annotate probable gene functions (*www.geneontology.org*). Because many gene functions are still unknown, complete information was often lacking on gene lists queried against the GO database. Specifically, the following steps were taken in generating entries in Table 1:

1. Column 1 represents a particular detected pattern type.
2. Column 2 represents the number of detected probe sets within each pattern type.
3. Column 3 filters entries from column 2 to remove what are broadly termed "probable contaminants." Each tumor sample harvested and arrayed does not contain purely tumor cells.

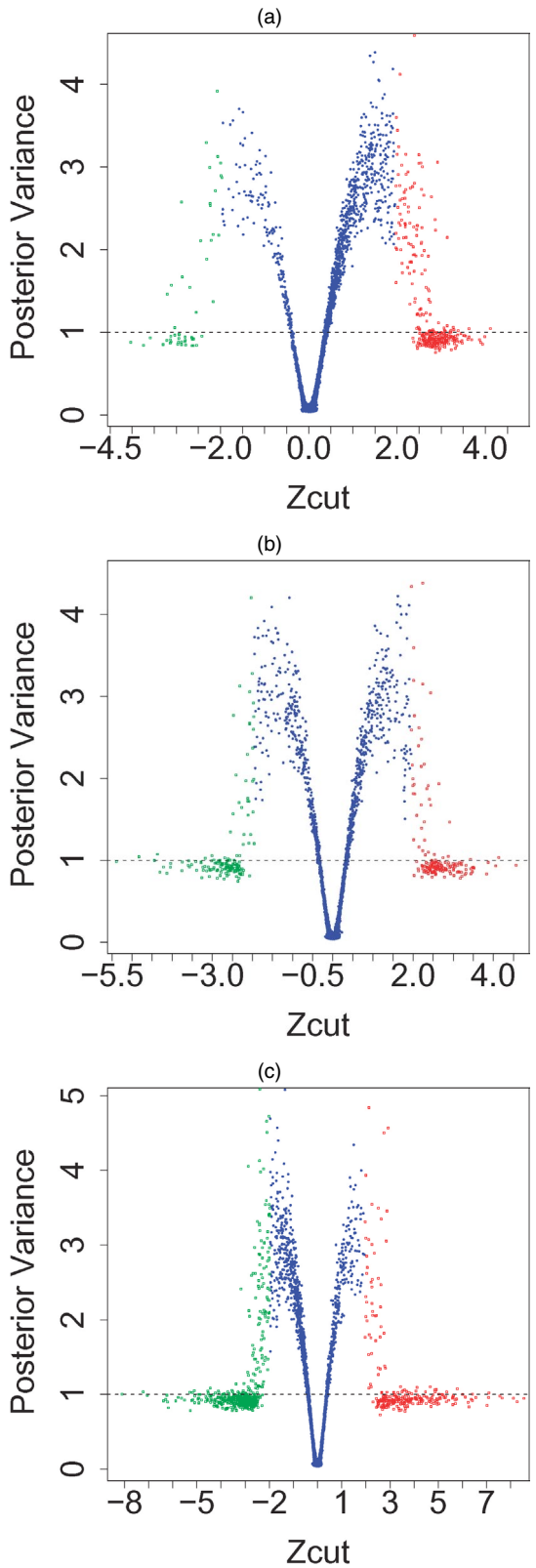

Figure 10. Shrinkage Plots. (a) C versus BSurvivors, 350 significant genes. (b) D versus BSurvivors, 338 significant genes. (c) METS versus BSurvivors, 774 significant genes. Color coding is defined as in Figure 5.

For instance, with a liver METS tissue sample, normal liver cells might compose a reasonable fraction of the sample. For earlier stages of colon cancer, immune cells involved in an inflammatory response might be present in the sample. We iden-

Table 1. Genes Mapped to Differential Expression Patterns for Multigroup Colon Cancer Analysis

Pattern	*Probe sets*	*Exclude*	*Remain*	*Genbank*	*Interesting pathways*
(1, 0, 0)	282	0	282	270	Cell cycle regulation,TGF-β signaling, cell adhesion, chromatin assembly
(0, 1, 0)	141	1	140	128	Apoptosis, MAPKKK cascade, transcription regulation, cell proliferation, oxidative stress response, cell adhesion, cell-to-cell signaling
(0, 0, 1)	258	80	178	172	Fatty acid synthesis/degredation, G-protein coupled receptor signalling, TGF-β signaling, blood clotting cascade, glycolysis/gluconeogenesis, matrix metalloproteinases
(1, 1, 0)	19	0	19	16	Development, transcription regulation
(0, 1, 1)	11	0	11	9	Transcription and translation regulation, cell-to-cell signaling
(1, 1, 1)	1	0	1	1	Unknown
(−1, 0, 0)	27	0	27	21	Oxidative stress response, GTPase activity
(0, −1, 0)	102	1	101	89	Tryptophan metabolism, galactose/phenylalanine metabolism, apoptosis
(0, 0, −1)	448	20	428	382	Cell cycle regulation, metabolism, Wnt signaling pathway, transcription regulation, G-protein coupled receptor signaling
(−1, −1, 0)	8	0	8	8	GTPase activity, proteolysis, muscle development
(0, −1, −1)	43	2	41	39	Apoptosis, immune response, frizzled signaling pathway, chemotaxis/inflammatory response
(−1, −1, −1)	13	3	10	7	Tryptophan assembly

NOTE: Columns are pattern type; number of probe sets found on gene chip with a pattern type; number of probe sets excluded; number of probe sets remaining after exclusion; number of probe sets with a Genbank Accession number; and interesting pathways found.

tified a list of genes that are known to express very highly in these contaminant cell types. Due to the mixture composition of the samples, if any of probe sets within a list under a pattern type mapped to a contaminant gene, they were filtered out because we could not be sure whether measured expressions were specific to tumor cells.

4. Column 4 reports the number of probe sets within each pattern type remaining after filtering.

5. Column 5 reports the number of remaining probe sets with Genbank Accession numbers. The Genbank Accession number assigns a numeric value to DNA sequences that have been studied to date.

6. The Genbank Accession numbers were then converted to gene symbols (*www.ncbi.nlm.nih.gov*). Gene symbols were then interrogated for biological functionality (known pathways of action) using the method of Liu et al. (2003) which produces GO annotations. Because genes are sometimes represented by more than one probe set on the gene chip, another filtering process was carried out, resulting in refined inference at the gene level only. To minimize the influence of falsely detected genes, only those pathways representing multiple times are highlighted in the table.

6.1 Hit-and-Run Genes

Clearly, many interesting patterns are listed in Table 1. For ease of discussion, we focus on three particular hit-and-run patterns: the (1, 0, 0), (0, 1, 0), and (0, 0, 1) genes. These indicate genes up-regulated with biological action specific to the various stages of disease only. The intent here is to try to *characterize* a particular stage genomically with respect to the baseline group BSurvivors. One can think of our analysis of hit-and-run genes (or, for that matter, of any other type of pattern) as a way to identify a group of genes with interesting activity relative to the BSurvivors, which can then be studied further for other types of interesting patterns. For instance, a gene in the (0, 1, 1) pattern type might in fact exhibit a significant difference from the colon cancer Duke's D to the liver METS—that is, this gene might exhibit continually increasing gene expression from the Duke's D to the liver METS. This can be ascertained by follow-up analyses of the (0, 1, 1) pattern.

Getting back to the hit-and-run patterns, it is helpful to mention some background on what is currently known about the biological pathways involved in the various stages of colon cancer (see Hegde et al. 2001; Nelson 1998; Quillin 2000; John 2001 for more details). Metastasis is broadly defined as the formation of secondary foci at a site distant from the primary site of origin. This process involves a series of interdependent, sequential events including initial growth, angiogenesis, invasion, extravasation, and establishment of new growth at the secondary site. These events involve cell cycle regulation, control of cell adhesion pathways, hypoxia resistance, and glycolysis/gluconeogenesis activation.

Specifically, progression of the cell cycle from its initial growth phase (G1) to its mitotic phase (M) is driven by positive and negative regulators that ultimately direct the fate of a cell either to form two daughter cells or to enter into resting state (G0). Dysregulated cellular proliferation, arising from

abnormal expression of genes that control cell cycle checkpoints (G1-S and G2-M phases)—that is, cell cycle pathways—are thus critical to the initial steps of tumorigenesis. Looking at genes in pattern $(1, 0, 0)$ that correspond to early changes (Duke's C) relative to a BSurvivor tumor, we see that cell cycle control and cell proliferation are definitely involved.

For a tumor to become invasive, it must pass through the muscularis mucosa and infiltrate the subserosal layer in which terminal lymphatics reside. Subsequently, genes involved in breaking the barriers of cellular adhesion play an important role in tumor invasiveness. This is manifested in pattern $(0, 1, 0)$, which are genes that have Duke's D stage—specific activity (relative to the BSurvivors). Recall that tumors in Duke's D stage are those left over as a deposit deep in the colon wall after the tumor has metastasized to the liver. Clearly, a significant amount of tumor invasion must have occurred for this to take place.

Pattern $(0, 0, 1)$ corresponds to genes up-regulated (relative to BSurvivors) specific to liver metastasis only. Note that many interesting pathways are listed, but in particular genes involved in glycolysis and gluconeogenesis were found. This is interesting because metastatic tumors will experience oxidative stress due to the fact that constituent cells are replicating rapidly (Dang and Semenza 1999). In effect, the tumor burns out its energy supply and has to create new energy sources de novo. To do this, new glucose synthesis pathways are tapped and glycolysis is instigated that metabolizes only glucose, not fats or other carbohydrates (John 2001). In fact, low-carbohydrate/low-glucose diets are coming into vogue as treatment options for aggressive tumors (Quillin 2000).

So using these three patterns only, genes mapped to known implicated biological pathways can be found. This opens up new potential therapeutic and diagnostic targets and demonstrates the power of the new methodology. Clearly, there are many other interesting biological findings. For instance, notice how only a single gene was found to be up-regulated throughout all of the stages of tumor development. Although this might be a false detection, it might also be a very significant finding. Unfortunately, this gene did not have any known biological function.

Remark 6. Comparing mirror image patterns that involve significant METS effects (i.e., patterns that differ only in sign for the METS), it appears from Table 1 that many more genes were down-regulated than up-regulated from our analysis. This is likely not the case, because we discovered some RNA degradation of the liver METS samples. This was due to some handling and preparation issues at the laboratory and likely inflated the number of genes mapped as down-regulated at the liver METS stage, but also potentially decreased the number of up-regulated genes found for these patterns.

7. DISCUSSION

One of the points made in this article is that spike and slab shrinkage pays dividends in terms of low false-detection rates while maintaining high power. Clearly, in multigroup data, large numbers of false detections over different group patterns cloud biological interpretation and can lead to wasted resources in downstream analyses. At the same time, power is important, because we would not want to miss an important effect. A rescaled spike and slab model is able to achieve these two opposing goals because of the effects of selective shrinkage. Selective shrinkage is accentuated in sparse settings, as was demonstrated theoretically. This phenomenon was also demonstrated at a practical level. In our colon cancer analysis, our new shrinkage estimates give very few sawtooth patterns, which are most likely biologically implausible. In addition, very few genes fell into the bottom-right and top-left quadrants of Figure 1. This represents a situation where the biological effect remained significant in sequential stages but was accompanied by a change in sign. It is true that this phenomenon is also at play for the usual Z-test estimates, but there protection against this type of finding is due mostly to the positive correlation between parameter estimates and is to be expected a priori. In general this correlation works against Z-test and will inflate false detection rates, as we saw with sawtooth genes. We found that the number of sawtooth pattern genes identified by Z-test was dramatically larger than that identified by Zcut, because of regression to the mean. As explained earlier, selective shrinkage greatly diminishes this kind of effect.

In this article we focused on looking for differentially expressing genes over all possible group patterns. This is the natural extension of the two-group problem to the multigroup setting and reflects the growing complexity in the type of data and questions that the applied scientist is seeking to answer through microarrays. Other patterns of interest might be genes exhibiting specific monotonic-increasing or -decreasing differential expression trends. If covariates other than group membership were available, then it would be of interest to study these same kinds of questions while adjusting for the additional predictors. One of the strengths of the rescaled spike and slab approach is that it ultimately rests on a flexible linear regression framework. Here this framework was recast as an ANOVA model to handle some of the issues arising in a stagewise analysis of colon cancer, but certainly the method could be modified and adjusted to handle other types of problems, such as those just outlined.

User-friendly, Java platform-independent software that implements the methods discussed here is available. Readers interested in this software should contact the authors by e-mail or visit their respective web pages.

APPENDIX: PROOFS

Proof of Theorem 1

Let $\mathbf{Z}_{n,j} = n_g^{1/2}(\overline{Y}_{1,j} - \theta_j, \ldots, \overline{Y}_{g,j} - \theta_j)^t$. A triangular central limit theorem shows that $\mathbf{Z}_{n,j} \overset{\mathrm{d}}{\rightsquigarrow} \mathrm{N}(\mathbf{0}, \sigma_0^2 \mathbf{\Pi})$ under the null, where $\mathbf{\Pi} = \mathrm{diag}\{\Pi_{g,0}/\Pi_{1,0}, \ldots, \Pi_{g,0}/\Pi_{g-1,0}, 1\}$. Observe that $\boldsymbol{\beta}_j^* = \widehat{\sigma}_N^{-1} \mathbf{\Sigma}_j \times \mathbf{B}_n \mathbf{Z}_{n,j}$, where

$$\mathbf{B}_n = \begin{pmatrix} b_{1,n} & 0 & 0 & \cdots & 0 & -b_{1,n} \\ 0 & b_{2,n} & 0 & \cdots & 0 & -b_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b_{g-1,n} & -b_{g-1,n} \end{pmatrix}_{(g-1)\times g}$$

and $b_{k,n} = (1 + n_g/n_k)^{-1/2}$. It can be shown that $\widehat{\sigma}_N^2 \overset{\mathrm{p}}{\to} \sigma_0^2$. Therefore, $\boldsymbol{\beta}_j^* \overset{\mathrm{d}}{\rightsquigarrow} \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_j \mathbf{B}_0 \mathbf{\Pi} \mathbf{B}_0^t \mathbf{\Sigma}_j^t)$, where $\mathbf{B}_0$ is the limit of $\mathbf{B}_n$ obtained by replacing $b_{k,n}$ by its limiting value $b_{k,0}$. The result follows by checking that $\mathbf{B}_0 \mathbf{\Pi} \mathbf{B}_0^t = \mathbf{\Omega}$.

Proof of Theorem 2

By definition, $\widehat{\beta}^*_{k,j} = V_{k,j}\widehat{\beta}_{k,j}$, where $V_{k,j} = \gamma_{k,j}/(\gamma_{k,j}+1)$. Under the assumption of normality, $\widehat{\sigma}_N\widehat{\beta}_{k,j}$ has a $\mathrm{N}(m_{k,j}, \sigma_0^2)$ distribution, where $m_{k,j} = \beta_{k,j,0}/\sqrt{1/n_k + 1/n_g}$. Let $\mathscr{I}_{j,0} = \{k : \beta_{k,j,0} = 0\}$ be the indices for the zero $\beta_{k,j,0}$ coefficients, and let $n_{j,0}$ be the cardinality of $\mathscr{I}_{j,0}$. By assumption, $n_{j,0} \geq 1$. Choose $\mathbf{\Gamma}_j$ such that $V_{k,j} = \alpha_1$ for each $k \in \mathscr{I}_{j,0}$ and $V_{k,j} = \alpha_2$ for each $k \in \mathscr{I}^c_{j,0}$, where $0 < \alpha_1, \alpha_2 < 1$ are values to be specified. Therefore,

$$\begin{aligned}
&\sum_{k=1}^{g-1} \mathbb{P}\{\widehat{\delta}_{k,j}(C) \neq \delta_{k,j}\} - \sum_{k=1}^{g-1} \mathbb{P}\{\widehat{\delta}^*_{k,j}(C) \neq \delta_{k,j}\} \\
&\quad = n_{j,0}\big(\mathbb{P}\big\{|\mathrm{N}(0,\sigma_0^2)| \geq C\widehat{\sigma}_N\big\} - \mathbb{P}\big\{|\mathrm{N}(0,\sigma_0^2)| \geq C\alpha_1^{-1}\widehat{\sigma}_N\big\}\big) \\
&\qquad + \sum_{k \in \mathscr{I}^c_{j,0}} \big(\mathbb{P}\big\{|\mathrm{N}(m_{k,j},\sigma_0^2)| < C\widehat{\sigma}_N\big\} \\
&\qquad\qquad - \mathbb{P}\big\{|\mathrm{N}(m_{k,j},\sigma_0^2)| < C\alpha_2^{-1}\widehat{\sigma}_N\big\}\big), \qquad \text{(A.1)}
\end{aligned}$$

where the $\mathbb{P}$-distributions on the right side correspond to the joint distribution for a normal random variable and the distribution for $\widehat{\sigma}_N$, where $\widehat{\sigma}_N^2/\sigma_0^2$ has an independent chi-squared distribution with $(n-g)M$ degrees of freedom. It is clear that the sum on the right side can be made arbitrarily close to 0, uniformly for $C_1 < C < C_2$, by choosing α_2 close to 1, whereas the first term on the right side remains positive and uniformly bounded away from 0 for any $\alpha_1 < 1$. Thus (A.1) is positive, uniformly over $C_1 < C < C_2$ if α_2 is chosen suitably.

Proof of Theorem 3

Let $\pi^*(\cdot|\mathbf{w})$ be the posterior measure for $\boldsymbol{\gamma}$ conditional on $\mathbf{w} = (w_1, \ldots, w_{g-1})$. Using conjugacy, and taking advantage of the orthogonality of $\mathbf{X}_j$, we can show that $\pi^*(\cdot|\mathbf{w})$ is defined by

$$\pi^*(B|\mathbf{w}_k) = \frac{\int_B \psi(\boldsymbol{\gamma})\pi(d\boldsymbol{\gamma}|\mathbf{w})}{\int \psi(\boldsymbol{\gamma})\pi(d\boldsymbol{\gamma}|\mathbf{w})} \qquad \text{(A.2)}$$

for each set B, where $\psi(\boldsymbol{\gamma}) = \prod_{t=1}^{M}\prod_{s=1}^{g-1}\psi_{s,t}(\gamma_{s,t})$ and

$$\psi_{s,t}(\gamma_{s,t}) = \exp(\nu_{s,t}\xi^2_{s,t})(1+\gamma_{s,t})^{-1/2}.$$

Note that $\gamma_{s,t}$ are conditionally independent given $\mathbf{w}$. Therefore,

$$\mathbb{E}^*(\nu_{k,j}|w_k) = \frac{\int \nu_{k,j}\psi_{k,j}(\gamma_{k,j})f_{k,j}(\gamma_{k,j}|w_k)\,d\gamma_{k,j}}{\int \psi_{k,j}(\gamma_{k,j})f_{k,j}(\gamma_{k,j}|w_k)\,d\gamma_{k,j}}.$$

Using (5), it is not difficult to show that $f_{k,j}(\cdot|w_k) = (1-w_k)g_0(\cdot) + w_k g_1(\cdot)$. Therefore, making the change of variables from $\gamma_{k,j}$ to $\nu_{k,j}$, deduce (12).

Proof of Corollary 1

Change variables in (12) from ν to $u = 1-\nu$. Multiply the numerator and denominator by $\exp(-\xi^2_{k,j})$ to get

$$\mathbb{E}^*(\nu_{k,j}|w_k) = 1 - \frac{\int_0^1 u^{-1/2}\exp(-u\xi^2_{k,j})f_{k,j}((1-u)/u|w_k)\,du}{\int_0^1 u^{-3/2}\exp(-u\xi^2_{k,j})f_{k,j}((1-u)/u|w_k)\,du}. \qquad \text{(A.3)}$$

To bound $\mathbb{E}^*(\nu_{k,j}|w_k)$, we will get an upper bound to the fraction on the right side. First, notice that for any $r \geq 0$,

$$u^r u^{-3/2} g_1\left(\frac{1-u}{u}\right) = \frac{D_1 u^{a_1+r-1/2}}{(1-u)^{a_1+1}}\exp\left(-\frac{a_2 u}{1-u}\right),$$

where $D_1 = a_2^{a_1}/(a_1-1)!$. Define

$$H(u) = (1-u)^{-(a_1+1)}\exp\left(-\frac{a_2 u}{1-u}\right).$$

It is clear that $H(u)$ is bounded above for $0 \leq u \leq 1$. Therefore, setting $r = 1$, and, using a similar argument to bound g_0, and hence $f_{k,j}$, deduce that

$$\begin{aligned}
&\int_0^1 u^{-1/2}\exp(-u\xi^2_{k,j})f_{k,j}\left(\frac{1-u}{u}\Big|w_k\right)du \\
&\quad \leq D_2\int_0^1 u^{a_1+1/2}\exp(-u\xi^2_{k,j})\,du \\
&\quad = D_2\xi_{k,j}^{-2(a_1+3/2)}\int_0^{\xi^2_{k,j}} u^{a_1+1/2}\exp(-u)\,du, \qquad \text{(A.4)}
\end{aligned}$$

where D_2 is a finite constant independent of w_k. Also,

$$H(u) \geq \exp\left(-\frac{a_2 u_0}{1-u_0}\right) \quad \text{for any } 0 \leq u \leq u_0,$$

where $u_0 < 1$ is some arbitrary fixed value. Therefore, setting $r = 0$, deduce that

$$\begin{aligned}
&\int_0^1 u^{-3/2}\exp(-u\xi^2_{k,j})f_{k,j}\left(\frac{1-u}{u}\Big|w_k\right)du \\
&\quad \geq D_3\int_0^{u_0} u^{a_1-1/2}\exp(-u\xi^2_{k,j})\,du \\
&\quad = D_3\xi_{k,j}^{-2(a_1+1/2)}\int_0^{u_0\xi^2_{k,j}} u^{a_1-1/2}\exp(-u)\,du, \qquad \text{(A.5)}
\end{aligned}$$

where D_3 is a finite constant independent of w_k.

If $\beta_{k,j,0} \neq 0$, then it can be shown that $\xi^2_{k,j} = O_p(n_k)$, using a triangular central theorem. Therefore, the integrals on the right side of (A.4) and (A.5) converge in probability to positive finite constants. Because the bounds (A.4) and (A.5) are independent of w_k, infer from (A.3) that

$$1 \geq \mathbb{E}^*(\nu_{k,j}) \geq 1 - O_p(\xi^{-2}_{k,j}) = 1 - O_p(n_k^{-1}).$$

Proof of Theorem 4

Recall that

$$\widehat{\beta}^*_{k,j} = \mathbb{E}^*(\nu_{k,j}) \times \frac{\overline{Y}_{k,j} - \overline{Y}_{g,j}}{\widehat{\sigma}_N\sqrt{1/n_k + 1/n_g}}.$$

First, we prove the necessity. Suppose that $|\widehat{\beta}^*_{k,j}| \xrightarrow{\mathrm{p}} \infty$. Because $0 \leq \mathbb{E}^*(\nu_{k,j}) \leq 1$ is bounded, it follows that $|\widehat{\beta}^*_{k,j}| \xrightarrow{\mathrm{p}} \infty$ if and only if $\beta_{k,j,0} \neq 0$. For sufficiency, suppose that $\beta_{k,j,0} \neq 0$. Then $\widehat{\beta}^*_{k,j} = O_p((1/n_k + 1/n_g)^{-1/2})$ and $|\widehat{\beta}^*_{k,j}| \xrightarrow{\mathrm{p}} \infty$. To complete the proof, we need to consider the behavior of $\mathrm{var}^*(\beta_{k,j})$. In particular, by Corollary 1 and (13), it suffices to show that $\mathrm{var}^*(\nu_{k,j}) = O_p(n_k^{-2})$.

Making a change of variables similar to (A.3) shows that

$$\begin{aligned}
&\mathbb{E}^*(\nu^2_{k,j}|w_k) \\
&\quad = 1 - 2A_n(w_k) + \frac{\int_0^1 u^{1/2}\exp(-u\xi^2_{k,j})f_{k,j}((1-u)/u|w_k)\,du}{\int_0^1 u^{-3/2}\exp(-u\xi^2_{k,j})f_{k,j}((1-u)/u|w_k)\,du},
\end{aligned}$$

where $A_n(w_k)$ is the fraction on the right side of (A.3). Define $B_n(w_k)$ to be the fraction on the extreme right side of the last expression. Then,

$$\begin{aligned}
\mathrm{var}^*(\nu_{k,j}) &= \mathbb{E}^*\big(\mathbb{E}^*(\nu^2_{k,j}|w_k) - \mathbb{E}^*(\nu_{k,j}|w_k)^2\big) + \mathrm{var}^*\big(\mathbb{E}^*(\nu_{k,j}|w_k)\big) \\
&= \mathbb{E}^*\big(1 - 2A_n(w_k) + B_n(w_k) - (1-A_n(w_k))^2\big) \\
&\quad + \mathrm{var}^*(A_n(w_k)) \\
&\leq \mathbb{E}^*\big(B_n(w_k) - A_n(w_k)^2\big) + \mathbb{E}^*(A_n(w_k)^2) \\
&= \mathbb{E}^*(B_n(w_k)) = O_p(n_k^{-2}),
\end{aligned}$$

where $B_n(w_k) = O_p(n_k^{-2})$, independently of w_k, is obtained using a similar proof as Corollary 1 (take $r = 2$).

[Received March 2004. Revised December 2004.]

REFERENCES

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.

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. DeVita, S. Hellman, and S. Rosenberg, Philadelphia, PA: Lippincott-Raven Publishers, pp. 1144–1196.

Dang, C. V., and Semenza, G. L. (1999), "Oncogenic Alterations of Metabolism," *Trends in Biochemical Science*, 24, 68–72.

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 Statical Association*, 96, 1151–1160.

Genovese, C., and Wasserman, L. (2002), "Operating Characteristics and Extensions of the FDR Procedure," *Journal of the Royal Statistical Society*, Ser. B, 64, 499–517.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Hegde, P., Qi, R., Gaspard, R., Abernathy, K., Dharap, S., Earle-Hughes, J., Gay, C., Nwokekeh, N. U., Chen, T., Saeed, A. I., Sharov, V., Lee, N. H., Yeatman, T. J., and Quackenbush, J. (2001), "Identification of Tumor Markers in Models of Human Colorectal Cancer Using a 19,200-Element Complementary DNA Microarray," *Cancer Research*, 61, 7792–7797.

Ishwaran, H., and Rao, J. S. (2003), "Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection," *Journal of the American Statistical Association*, 98, 438–455.

——— (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *The Annals of Statistics*, 33, 730–773.

John, A. P. (2001), "Dysfunctional Mitochondria, Not Oxygen Insufficiency, Cause Cancer Cells to Produce Inordinate Amounts of Lactic Acid: The Impact of This on the Treatment of Cancer," *Medical Hypotheses*, 57, 429–431.

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

Liu, G., Loraine, A. E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D., and Siani-Rose, M. A. (2003), "NetAffix: Affymetrix Probesets and Annotations," *Nucleic Acids Research*, 31, 82–86.

Nelson, N. J. (1998), "Quest for New and Better Colon Cancer Treatments Picks Up Steam," *Journal of the National Cancer Institute*, 90, 1858–1859.

Nguyen, D., Arpat, A. B., Wang, N., and Carroll, R. J. (2002), "DNA Microarray Experiments: Biological and Technological Aspects," *Biometrics*, 58, 701–717.

Quillin, P. (2000), "Cancer's Sweet Tooth," *Nutrition Science News*, 4, 1–8.

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

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 Science*, 98, 5116–5121.