Statistics and Its Interface Volume 3 (2010) 477–491

Variance model selection with application to joint analysis of multiple microarray datasets under false discovery rate control*

Long Qu, Dan Nettleton,† Jack C. M. Dekkers and Nicola Bacciu

We study the problem of selecting homogeneous variance models *vs.* heterogeneous variance models in the context of joint analysis of multiple microarray datasets. We provide a modified multiresponse permutation procedure (MRPP), modified cross-validation procedures, and the right AICc (corrected Akaike's information criterion) for choosing a variance model. In a simple univariate setting, our modified MRPP outperforms commonly used competitors. For microarray data analysis, we suggest using the sum of gene-specific selection criteria to choose one best gene-specific model for use with all genes. Through realistic simulations based on three real microarray studies, we evaluated the proposed methods and found that using the correct model does not necessarily provide the best separation between differentially and equivalently expressed genes, but it does control false discovery rates (FDR) at desired levels. A hybrid procedure to decouple FDR control and differential expression detection is recommended.

AMS 2000 subject classifications: Primary 62J20, 62F07; secondary 62P10, 92C40.
Keywords and phrases: AIC, AICc, Cross-validation, False discovery rates, Microarray, Model selection, Multiresponse permutation procedure, Variance model.

1. INTRODUCTION

1.1 Joint analysis of from multiple microarray datasets

Microarrays are a popular tool in genomic expression profiling studies for discovering genes that respond to treatments of interest. The measurements from each experimental unit in such studies are very high dimensional expression vectors, ranging from thousands to tens or hundreds of thousands of genes, far exceeding the available sample sizes. This creates a hurdle to analyzing the data from all genes simultaneously using traditional multivariate methods.

A popular and practical choice is to fit a univariate model, *e.g.*, a linear model or a linear mixed model [45], to each gene separately. For each contrast of interest, the analysis usually results in a p-value for each gene. This set of p-values is then summarized so that a certain error rate is controlled at a pre-specified level. False discovery rate (FDR) and variants thereof have become the *de facto* standard choices for this purpose [7, 40] due to the exploratory nature of microarray expression studies.

A major problem of this approach is that the power for detecting interesting genes is usually very low because 1) the cost of microarrays hinders most researchers from using moderate to large sample sizes, 2) the measurements produced by the current technology are rather noisy, and 3) the pre-existing biological variation among experimental units is often large. Although it may be reasonable to expect sample sizes to increase as the cost of technology decreases, an examination of the microarray datasets hosted in the Gene Expression Omnibus (GEO) database [5] indicates that the percentage of datasets with 10 or fewer experimental units has remained relatively steady at above 40% nearly every year over the past decade. Combining data from multiple datasets collected by a single lab or by several labs offers one way to address the persistent problem of insufficient sample size.

In this paper, we explore the variance model selection problem for joint analysis of multiple microarray datasets to improve the detection of differentially expressed genes using gene-wise linear models. In particular, we focus on how to determine if gene-specific estimates of error variance should be pooled across datasets and used for testing linear contrasts of means *within* a dataset. We assume that we have data from several similar datasets using the same type of microarray and biological samples from the same or very similar populations. Our main interest lies in testing linear contrasts *within* some datasets. This differs from the usual meta-analysis in that different datasets do *not* necessarily involve the same sets of treatments. Multiple datasets are used only to provide better variance estimates, instead of better estimates of means.

*The authors wish to acknowledge support from USDA-CSREES grants NRI-2005-3560415618 and 2008-35600-18786, and NSF grants DMS-0714978 and CCF-0811804. The authors are also grateful to the two reviewers and Dr. Philip Dixon for constructive comments.
†Corresponding author.

More specifically, the problem we consider can be described as follows. Suppose we have K independent data matrices, $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_K$, which correspond to K independent sets of high-dimensional vector-valued observations. We will refer to $\mathbf{Y}_k$ as a dataset and assume that $\mathbf{Y}_k$ is of dimension $G \times n_k$, with rows corresponding to genes and independent columns corresponding to vector-valued observations for all $k = 1, 2, \ldots, K$. Further, we assume that $\mathrm{E}(\mathbf{Y}_k) = \mathbf{B}_k'\mathbf{X}_k'$, where $\mathbf{X}_k$ is an $n_k \times p_k$ design matrix with full column rank and $\mathbf{B}_k$ is a $p_k \times G$ matrix of unknown parameters. For example, suppose the kth dataset follows a commonly used (unpaired) two-treatment design with 3 biological samples for each treatment. $\mathbf{X}_k$ could be $\mathbf{I}_2 \otimes \mathbf{1}_3$, where $\mathbf{I}_2$ is a 2×2 identity matrix, $\otimes$ stands for Kronecker product, and $\mathbf{1}_3$ is a vector of 3 ones. The corresponding $\mathbf{B}_k$ will be two rows of unknown mean parameters. Each row represents a treatment, and each column represents a gene. Finally, we assume that the covariance matrix of any column of $\mathbf{Y}_k$ is $\mathbf{\Sigma}_k$ for all $k = 1, 2, \ldots, K$. In this paper, we develop procedures for assessing whether a model assuming $\mathrm{diag}(\mathbf{\Sigma}_1) = \mathrm{diag}(\mathbf{\Sigma}_2) = \cdots = \mathrm{diag}(\mathbf{\Sigma}_K)$ should be selected for making inference about $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_K$.

1.2 Model selection

The aforementioned question is a model selection problem with respect to variances. In general, model selection has long been studied, and there are at least three major approaches. The first is to perform hypothesis testing, *e.g.*, a likelihood ratio test, and to choose a relatively simple model as long as there is no obvious evidence showing that the more complicated alternative models fit the data much better. In our problem, we could begin by testing for equality of variance across all K datasets within all genes. If the null hypothesis is not rejected, then pooling variance estimates across all datasets within each gene might be a good choice. If the null hypothesis is rejected, then a refined form of the variance model could be formulated and tested. For example, if one dataset seems to have increased variation relative to the rest, a model that allows the variances for that dataset to differ from the variances for other datasets could be proposed and tested. This procedure can be repeated until a sufficiently rich variance model is judged to adequately fit the data according to the testing procedure. The second approach is to order the candidate models according to some criterion and to choose the model that optimizes the criterion, *e.g.*, Akaike's information criterion (AIC) [1, 2], Schwarz Bayesian information criterion (BIC) [32], or prediction sums of squares (PRESS) [3]. The third approach is to employ Bayesian techniques and to choose models by summarizing posterior distributions [38]. Because Bayesian methods often rely heavily upon Monte Carlo simulations that are very computationally intensive for large microarray datasets, we only discuss the first two approaches in this study.

In the context of choosing an appropriate variance model for joint analysis of multiple datasets, the hypothesis testing approach requires a good test for heterogeneity in variances. For univariate data, the likelihood ratio tests, *i.e.*, the F-test for two-sample comparison and Bartlett's test [6, 37] for general one-way designs, are known to be sensitive to departures from normality. If the normality assumption holds, they have very good theoretical properties, but in practice, robust tests for variances are often recommended. For example, Levene's test performs one-way ANOVA on absolute residuals from a least squares fit [24]. The Brown-Forsythe test is similar but computes residuals from a least absolute deviations fit [9]. These tests are only approximate, and, as will be seen in our simulations, the approximation is often very poor under typical small sample sizes in microarray datasets. Thus, the resulting nominal p-value distribution deviates far from the theoretical uniform distribution on the unit interval. In this study, we propose an alternative permutation based procedure that better controls type I error and has good power. Moreover, it can be automatically applied to any high-dimensional dataset.

Commonly used model selection criteria can also be classified into three categories. Methods in the first category seek models that minimize some estimate of the prediction sums of squares. In linear models, Mallows' C_P [26] and PRESS are widely used methods for such a purpose. However, these methods are designed for selecting mean structures, *i.e.*, regressors of the model. In this study, we propose two new cross-validation measures designed specifically for differentiating alternative variance models.

The second category of model selection criteria, exemplified by AIC, corrected AIC (AICc) [21], and Takeuchi's information criterion (TIC) [41], includes methods that approximate expected estimated Kullback-Leibler divergence as a criterion to rank candidate models. These methods have firm information-theoretic justifications [1, 2] and are known to be asymptotically efficient [33, 34]. They do not intend to choose a smallest true model asymptotically but to choose a good approximate model based on the available amount of data, because the true model may be infinite dimensional and fall outside the set of candidate models. Among the three methods, AIC is a special case of TIC, but TIC is difficult to estimate and is rarely used. The AICc is a bias-corrected version of AIC, but its derivation is model dependent. Unfortunately, common software implementations often ignore this fact and compute a panacea version of AICc that assumes homogeneity in variances. For example, as of version 9.2 of `SAS/STAT` [31], neither the `PROC MIXED` procedure nor the `PROC GLIMMIX` procedure reports the correct AICc when a `GROUP` option is used in the `REPEATED` or `RANDOM` statement to specify heterogeneity of variances. Hence, in this study, we will provide the correct AICc formula for linear models with heterogeneous variances.

The third category of model selection criteria includes model dimension consistent criteria, *e.g.*, BIC, Hannan and

Quinn's information criterion (HQIC) [17] and Bozdogan's consistent AIC (CAIC) [8]. These methods have the property that when the true models are indeed in the set of candidate models, then as sample size increases to infinity, a true model with the smallest model dimension will be selected. However, when the candidate set does not include any true models, these criteria asymptotically choose a small approximate model based on Kullback-Leibler divergence. For the subtle difference compared with the second category of criteria, see [10].

A completely different strategy to model selection is regularizing parameter estimates through various shrinkage methods. In the context of microarray data analysis, many such methods have been proposed in the literature to combat the insufficient sample size problem in estimating variances. Most of these methods do not require the use of datasets from other studies. They aim to improve gene-wise variance estimation by pooling information from other genes within a single dataset. Examples of such methods are *ad hoc* modifications [14, 44], estimates based on mean-variance relationship [13, 15, 20, 22, 23], and hierarchical model based estimates [4, 11, 25, 36]. In this paper, we pick the very popular *limma* method [36] as a representative of such shrinkage estimation methods and compare its performance with our model selection approaches.

2. METHODS

In subsections 2.1 through 2.3, we propose three independent variance model selection approaches. We develop an approximate permutation test for testing homogeneity of high dimensional spread in subsection 2.1. In subsection 2.2, we propose cross-validation methods for choosing between variance models. Under normal theory linear model assumptions, the correct AICc formula for selection of the variance model is derived in subsection 2.3. Together with other information criteria, we suggest in subsection 2.4 to use the sum of information criteria across genes as a means to select a common variance model for all genes in the analysis of multiple microarray datasets. The performance of these methods is assessed through simulations based on real microarray data in subsection 2.5 and section 3.

2.1 Modified MRPP for testing homogeneity of variances

The multi-response permutation procedure (MRPP) [27] is a multivariate permutation test that has been successfully applied to the analysis of gene sets for microarray data [28], with the advantage that the dimension of the response variable needs not be less than the sample size. For a K-treatment design, the usual MRPP statistic is constructed in two steps. First, average pairwise Euclidean distances across observations within each treatment group are calculated as a measure of spread within the treatment group. Next, the test statistic is constructed by a weighted sum of the K average pairwise distances, which is largely motivated by the decomposition of sums of squares in the usual analysis of variance (ANOVA). The treatment labels are then randomly shuffled a large number of times, and the test statistic is re-computed for each shuffling. The p-value is reported as the proportion of shufflings that result in a test statistic no larger than the one observed for the original data before shuffling.

There are two obstacles in directly applying the usual MRPP to test the equality of variances across multiple microarray datasets. First, the usual MRPP test statistic is rather insensitive to changes in spread of the multivariate distribution because it is mainly designed for detecting mean differences, as in ANOVA. Second, within each microarray dataset, observations are not exchangeable due to the differences in means across treatments within datasets and the MRPP statistic does not account for this more complicated mean structure.

To overcome these problems, we first conduct a complete QR decomposition for each $\mathbf{X}_k$,

$$\mathbf{X}_k = \mathbf{Q}_k\mathbf{R}_k = [\mathbf{Q}_{k1}, \mathbf{Q}_{k2}][\mathbf{R}'_{k1}, \mathbf{0}]',$$

where $\mathbf{Q}_k$ is an $n_k \times n_k$ orthonormal matrix with the first p_k columns being $\mathbf{Q}_{k1}$ and the remaining columns being $\mathbf{Q}_{k2}$, and $\mathbf{R}_k$ is an $n_k \times p_k$ matrix with the first p_k rows being an upper triangular matrix $\mathbf{R}_{k1}$ and the remaining rows being zero.

To remove the mean structure in $\mathbf{Y}_k$, it is trivial to check that the orthogonal projection matrix (the hat matrix) that projects onto the column space of $\mathbf{X}_k$ is $\mathbf{Q}_{k1}\mathbf{Q}'_{k1}$, so that the residual matrix $\mathbf{Y}_k(\mathbf{I}_k - \mathbf{Q}_{k1}\mathbf{Q}'_{k1})$ has mean zero, where $\mathbf{I}_k$ is the $n_k \times n_k$ identity matrix. Mielke and Berry [27] directly used the columns of residual matrices similar to these to perform permutation tests. However, this is problematic because 1) the columns within the same residual matrix are correlated, whereas columns from different residual matrices are independent, and 2) the covariance matrices corresponding to columns within any residual matrix are not necessarily identical. Hence, even if the null hypothesis is true, the fundamental assumption of exchangeability for the permutation test is violated.

To improve exchangeability, we suggest using a transformation that simultaneously removes the means, decorrelates the columns, and standardizes the columns. Specifically, we define $\mathbf{Z}_k = \mathbf{Y}_k\mathbf{Q}_{k2}$ for all $k = 1, 2, \ldots, K$ and propose to use the columns of $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_K$ as data for an MRPP-based (approximate) test of variance equality. Note that each $\mathbf{Z}_k$ is $G \times d_k$ instead of $G \times n_k$, where $d_k = n_k - p_k$ is the error degrees of freedom for dataset k. This reduction makes sense since after accounting for the means, the effective sample size for the variances is actually d_k for the kth dataset. A similar reduction is used for residual maximum likelihood [29] estimation of variance components in mixed linear models. Tests not accounting for the loss of

information in estimating the location parameters, *e.g.*, the commonly used Levene's test and the Brown-Forsythe test, will likely lead to poorly controlled type I errors, as verified through simulations in the next section. Also, by using the orthogonality of the columns of $\mathbf{Q}_k$, it can be easily checked that the columns of $\mathbf{Z}_k$ are uncorrelated and the covariance matrix of any column of $\mathbf{Z}_k$ is identically $\mathbf{\Sigma}_k$, the same as the covariance matrix of any column in $\mathbf{Y}_k$. Although uncorrelatedness implies neither independence nor exchangeability among the columns, unless normality is further assumed, $\mathbf{Z}_k$ does offer a better candidate than the raw residuals for approximate permutation tests, by alleviating the problem up to the second order moments.

Lastly, we need a test statistic based on these $\mathbf{Z}$ matrices. Both steps in constructing the usual MRPP statistic need some modification. First, the measure of within-treatment spread no longer needs the "pairwise" concept, because we now know the location of $\mathbf{Z}_k$, *i.e.*, $\mathrm{E}(\mathbf{Z}_k) = \mathbf{0}$. Actively using this location information will lead to more powerful tests. To keep our approach closely related to the well established MRPP statistic, we choose the spread measure to be the average Euclidean distance of each column of $\mathbf{Z}_k$ to the *origin* and denote this as δ_k. Second, for testing the equality of spreads across the K datasets, the weighted sum of δ_k used in the usual MRPP formulation is no longer sensible. Instead, we propose to use

$$\frac{\min\limits_{k=1,2,\ldots,K} \delta_k}{\max\limits_{k=1,2,\ldots,K} \delta_k}$$

as the test statistic, which is similar to the $F_{\max}$ statistic of Hartley [19]. Analogous to standard permutation procedures, the dataset labels for columns of $\mathbf{Z}_k$'s are then shuffled a large number of times and the p-value is reported as the proportion of shufflings that results in a test statistic smaller or equal to the one observed without shuffling.

This test will be more sensitive when only one or a few of the K datasets have very different spreads. Other statistics can be constructed to be more sensitive when most of the K datasets have slightly different spreads, but we believe the latter case is less important for the purpose of joint analysis of multiple datasets because most tests of mean contrasts can tolerate mild departures from the equal variance assumption, especially under balanced designs. Second, this statistic is more sensitive to differences in marginal variability than to changes in correlations among the G dimensions. This is advantageous for joint analysis of multiple microarray datasets because, in the standard gene-by-gene analysis, pooling error variance estimates across datasets is justified whenever marginal variances are approximately constant across datasets. Thus for the microarray application, it is more important to be able to detect departures from $\mathrm{diag}(\mathbf{\Sigma}_1) = \mathrm{diag}(\mathbf{\Sigma}_2) = \cdots = \mathrm{diag}(\mathbf{\Sigma}_K)$ than to be able to detect departures from $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_K$. Furthermore, our experience with real microarray datasets suggests that an expansion of gene-specific error variances by a multiplicative factor from one dataset to another is the most common type of departure from $\mathrm{diag}(\mathbf{\Sigma}_1) = \mathrm{diag}(\mathbf{\Sigma}_2) = \cdots = \mathrm{diag}(\mathbf{\Sigma}_K)$. Thus, the power of our test is focused on detecting such alternatives.

2.2 Cross-validation for selecting variance models

Cross-validation (CV), as a data based method for estimating prediction ability, is a powerful tool for model selection. However, the most commonly used CV method in linear models, PRESS, and the closely related Mallows' C_P criterion, are not able to identify heterogeneity in variance. Furthermore, the usual application of CV only selects one common model that has good prediction ability over all K datasets. But when our interest primarily lies in inference *within* a single dataset, the selected model might not be optimal because the prediction ability on other datasets biases our choice. Hence, it might be more interesting to select different models for inference problems within different datasets.

Here we propose CV based procedures that can solve the above problems. A key point is to consider element-wise squared $\mathbf{Z}_k$, denoted as $\mathbf{Z}_k^{(2)}$, as the raw data to perform prediction, such that the selection of variance models on $\mathbf{Z}_k$ can be roughly treated as the selection of mean models on $\mathbf{Z}_k^{(2)}$. For ease of discussion, let $\mathbf{z}_k^{(2)}$ denote an arbitrary row of $\mathbf{Z}_k^{(2)}$. The same procedure to be discussed can be applied to each row of $\mathbf{Z}_k^{(2)}$. In the case of normality based linear models for $\mathbf{Y}_k$, the elements of $\mathbf{z}_k^{(2)}$ are independently and identically distributed as scaled χ_1^2 random variables with mean $\sigma_k^2 > 0$. Hence the average of all elements of $\mathbf{z}_k^{(2)}$ provides a natural estimator of its mean, *i.e.*, the variance of each element in the corresponding row in $\mathbf{Y}_k$. A variance model can be specified by a function M that maps dataset $k \in \{1, 2, \ldots, K\}$ *onto* $\{1, 2, \ldots, J\}$, such that $\sigma_k^2 = \sigma_{k'}^2$ iff $M(k) = M(k')$ for $J \in \{1, 2, \ldots, K\}$. For a full model, $J = K$; and for a reduced model, $J < K$. Once the $\mathbf{z}_1^{(2)}, \mathbf{z}_2^{(2)}, \ldots, \mathbf{z}_K^{(2)}$ are treated as the data, leave-one-out cross-validation can be done as usual. That is, we delete one data point and use the average of other data points that share the same mean according to the model specified by M as a predictor for the deleted data point, and the procedure is repeated for each data point.

Suppose the ith element of $\mathbf{z}_k^{(2)}$, $z_{ki}^{(2)}$, is deleted, and the prediction based on the remaining data according to the model specified by M for $z_{ki}^{(2)}$ is

$$\hat{z}_{k(-i)}^{(2)} = \frac{1}{D_k - 1}\left(-z_{ki}^{(2)} + \sum_{\{k':M(k')=M(k)\}} \sum_{j=1}^{d_{k'}} z_{k'j}^{(2)}\right),$$

where $D_k = \sum_{\{k':M(k')=M(k)\}} d_{k'}$ is the total degrees of freedom used for estimating the mean of $z_{ki}^{(2)}$ according to model

M *without* deleting any data. We define the raw PRESS residual as $e_{ki} = z^{(2)}_{ki} - \tilde{z}^{(2)}_{k(-i)}$ and the corrected PRESS residual as $e'_{ki} = e_{ki}\sqrt{(D_k - 1)/(2D_k)}$. It can be easily checked that $\mathrm{E}(e_{ki}) = \mathrm{E}(e'_{ki}) = 0$ and $\mathrm{Var}(e'_{ki}) = \sigma^4_k$. The multiplicative correction in e'_{ki} removes the dependence of $\mathrm{Var}(e_{ki})$ upon D_k, so that e'_{ki} are comparable to each other. Although this correction assumes normality of $\mathbf{Y}_k$, it is still reasonable in other cases, especially for reasonable sample sizes. The final PRESS measure for model selection is then $\sum_{k=1}^{K}\sum_{i=1}^{d_k} {e'_{ki}}^2$. The model with the smallest PRESS measure is preferred.

Alternative to the squared error loss above, we may choose the prediction loss to be the $(-2\times)$ prediction log likelihood for each $z^{(2)}_{ki}$, which relates to the Kullback-Leibler divergence. Let the log likelihood based on a single data point $z^{(2)}_{ki}$ be $\ell_{ki}(\sigma^2_k) = \log f_{\chi^2_1}(z^{(2)}_{ki}/\sigma^2_k) - \log\sigma^2_k$, where $f_{\chi^2_1}$ is the density function for the χ^2_1 distribution. The final CV based prediction log likelihood criterion is then $-2\sum_{k=1}^{K}\sum_{i=1}^{d_k}\ell_{ki}(\tilde{z}^{(2)}_{k(-i)})$. Again, the model with the smallest criterion is preferred.

Note that the above two procedures select one model to be used for all K datasets. This may not be ideal if we are interested in analyses within datasets but not across datasets. In this case, CV based procedures can be easily adapted to dataset-specific PRESS $\sum_{i=1}^{d_k}{e'_{ki}}^2$ and/or dataset-specific $(-2\times)$ prediction log likelihood $-2\sum_{i=1}^{d_k}\ell_{ki}(\tilde{z}^{(2)}_{k(-i)})$ as model selection criteria. Note that even if these procedures are dataset-specific, information from other datasets is borrowed through the prediction $\tilde{z}^{(2)}_{k(-i)}$. Use of these dataset-specific criteria would allow different variance pooling strategies for different datasets. As we will demonstrate through simulation in the next section, in some cases it may be advantageous to pool error variance estimates across two datasets 1 and 2 for analyzing dataset 1 but not for analyzing dataset 2. Thus such a breakdown of model selection criteria can be quite useful. Note that it is rarely straightforward to break other information criteria (*e.g.*, AIC or BIC) down in this manner, because of the difficulty in reasonably decomposing the penalty on log likelihood.

2.3 AICc under heteroscedasticity

Hurvich and Tsai [21] developed AICc for selecting an appropriate mean model in a linear model context under the normality assumption. As an estimate of expected estimated Kullback-Leibler divergence, AICc is exactly unbiased and has the same variance as the asymptotically unbiased AIC. Hence Burnham and Anderson [10] recommended routine use of AICc over AIC.

However, such analytic small sample bias correction has to be dealt with case by case, *i.e.*, the correction will be different for different models. For example, the correction under a heterogeneous variance assumption is different than under a homogeneous variance assumption. Unfortunately, to our knowledge, implementations in common statistical software do not acknowledge such differences.

Under the fully heteroscedastic model across K datasets, the correct AICc formula is

$$-2\sum_{k=1}^{K}(\log \mathrm{REML}_k) + 2\sum_{k=1}^{K}\frac{d_k}{d_k - 2},$$

where REML_k is the maximized residual likelihood for dataset k. To see why this formula makes sense, first consider the kth dataset alone, where the homoscedastic AICc is known to be $-2\log\mathrm{REML}_k + 2d_k/(d_k - 2)$. That is, the bias in using maximized residual log likelihood to estimate the expected mean prediction log likelihood for the kth dataset is exactly $d_k/(d_k - 2)$. Since the K datasets are independent, their log likelihoods are additive and the final bias is simply the sum of the bias of each individual maximized residual log likelihood, which justifies the heteroscedastic AICc formula.

Because $2d_k/(d_k - 2)$ is always greater than 2 when $d_k > 2$, the AICc penalty is larger than the AIC penalty $2K$. When $d_k \to \infty\ \forall\ k$, the two criteria are equivalent. However, the `SAS/STAT PROC MIXED` procedure with the `GROUP` option for specifying heterogeneous variance across datasets uses $2K(\sum_{k=1}^{K} d_k)[(\sum_{k=1}^{K} d_k) - K - 1]^{-1}$ as the penalty term, which converges to AIC if $d_k \to \infty$ for *any* k. Hence we would expect that this incorrect AICc will not perform well when some d_k's are small but others are large.

2.4 Combining model selection criteria from multiple genes

In a usual gene-by-gene analysis of microarray data, one can compute information criteria for each gene separately. For better interpretability, it is often desirable to fit the same model to each gene, and hence an overall criterion for choosing one model based on data from all genes is needed.

We suggest using the sum of gene specific information criteria as an overall measure for model selection. This is a sensible strategy in general and the obvious strategy for the special case of independence across genes. For AIC, the log likelihood is additive under independence and the penalty term (the number of parameters) is also additive. So the sum of gene specific AIC's equals the AIC directly calculated by assuming an independence model on all G genes. The same argument also applies to prediction log likelihood CV and the correct AICc developed in the last subsection, but not for the AICc reported by `SAS/STAT`, nor for BIC, CAIC, or HQIC since their penalty terms cannot be added directly.

The Akaike's weights [10] can be used to justify the use of sums as an overall model selection criterion for BIC, CAIC, and HQIC, in addition to AIC and AICc. Considering q candidate models, the Akaike's weight for model m and gene g is defined as

$$w_{mg} = \frac{\exp\{-0.5\Delta\mathrm{IC}_{mg}\}}{\sum_{m'=1}^{q}\exp\{-0.5\Delta\mathrm{IC}_{m'g}\}},$$

where $\Delta\text{IC}_{mg} = \text{IC}_{mg} - \min_{1\le m'\le q}\text{IC}_{m'g}$, and IC_{mg} is the information criterion for model m and gene g, $m = 1, 2, \ldots, q$, and $g = 1, 2, \ldots, G$. This weight can be easily interpreted as the posterior probability of model m for gene g, and the choice of different information criteria is equivalent to choosing a different prior probability for each model (see [10] for details).

Hence, by assuming independence across genes, the posterior probability of selecting model m for all G genes is the product $w_{m\cdot} = \prod_{g=1}^{G} w_{mg}$, which is a monotonically decreasing function of the sum of information criteria $\sum_{g=1}^{G}\text{IC}_{mg}$. In other words, the model selected with the smallest sum of information criteria across genes always matches the model with the largest posterior probability among the $q-1$ alternative overall models.

To combine gene-specific PRESS into an overall model selection measure, the sum is not as reasonable because genes with large variances tend to have large PRESS and the sum may be dominated by such genes. Hence, we use gene voting to select the overall model, so that the model selected for all genes will be the one selected by the plurality of gene-specific PRESS statistics.

Although the above arguments rely on independence across genes, these combination procedures are still natural intuitive measures under dependence. We do not expect that dependence will seriously bias the selection of models. Taking the most extreme case as an example, suppose we only have two genes which are nearly completely dependent. Then the gene-wise model selection criteria will almost always select the same model, and the combination of criteria using either the sum or the votes will also select the same model. Hence the combination does not favor either simpler or more complex models, even in this highly dependent case. Therefore, these combination procedures remain sensible and useful in practice. Furthermore, although our approach was developed under the assumption of independence across genes, note that we evaluate its performance using a simulation procedure that includes correlation across genes as described in the next subsection.

2.5 Simulation based on real microarray data

Simulation studies to evaluate the performance of model selection methods often generate data from parametric models. Although such results are helpful to aid understanding the pros and cons of each method, they provide little information about actual performance on real datasets, especially when we have few clues about the general dependence structure among a large number of genes. To overcome such difficulties, we replace traditional model-based simulation with subsampling from an actual microarray dataset that involves at least one large treatment group. Treating the large treatment group as a population, we can simulate various multiple-dataset scenarios by drawing subsamples from the population and perturbing the data as described below.

Table 1. Simulation setting for the ratio of variances

μ_r	σ_r^2	$\text{E}(r_g)$
0	0	1 (null)
−0.057800	0.342^2	1
1.040812	0.342^2	3
1.551638	0.342^2	5
2.244785	0.342^2	10

Details about the populations actually used in our study are provided at the end of this subsection.

Let G denote the number of genes and n the number of biological samples in the population. Let $\mathbf{E}$ denote a $G \times n$ residual matrix obtained by subtracting the gene-specific mean log-scale expression value from the log-scale expression values of each gene. Because the population size n is large, we ignored the small dependence across the columns of $\mathbf{E}$. Also, we computed the sample standard deviation for each gene g as σ_{0g}, and treated these as known values due to the large number of degrees of freedom.

To generate two datasets, each involving two treatments, we first randomly sampled $2n_1 + 2n_2$ columns from $\mathbf{E}$ without replacement to form a $G \times (2n_1 + 2n_2)$ matrix, with the first $2n_1$ columns $\mathbf{Y}_1$ denoted as dataset 1 and the other $2n_2$ columns $\mathbf{Y}_2$ denoted as dataset 2. Within each dataset, a balanced 2-treatment comparison design was used. Sample sizes n_1 and n_2 were set to 3, 5, or 7, which are similar to those used in many microarray studies.

Next, we independently generated G random variables r_g, $g = 1, 2, \ldots, G$, from a log Normal(μ_r, σ_r^2) distribution, where the settings of μ_r and σ_r^2 are listed in Table 1. Then the gth row of $\mathbf{Y}_2$ was multiplied by $\sqrt{r_g}$ for all $g = 1, 2, \ldots, G$. Hence r_g is the ratio of the error variance in the 2nd dataset to the error variance in the 1st dataset for the gth gene. Note that, when $\mu_r = \sigma_r^2 = 0$, all $r_g = 1$, *i.e.*, the error variances are equal across the two datasets. When $\sigma_r^2 > 0$, μ_r was chosen such that $\text{E}(r_g)$ was 1, 3, 5, or 10 respectively. $\sigma_r^2 = 0.342^2$ was used here such that the error variance correlation between the two datasets on the log scale is about 0.8, which is very similar to the observed log variance correlations between different treatments in the real microarray datasets we examined. In other words, our simulation setup acknowledges that genes with small error variances in one dataset also tend to have small variances in the other dataset, which is also a biologically rational assumption.

In order to assess the ability of different model selection methods to detect differentially expressed genes and to control false discovery rates, we perturbed the mean of a subset of genes in each dataset to mimic responses to the treatment. Since we believe that in real biological systems the standardized effective sizes are probably not related to the error variances, we added to each row g of the first treatment group in the kth dataset $\sigma_{0g} r_g^{(k)} u_{gk}$, where u_{gk} was independently sampled from a mixture of zero with probability 0.8

and a standard normal distribution with probability 0.2 for all $g = 1, 2, \ldots, G$ and $k = 1, 2$, and $r_g^{(k)} = 1$ when $k = 1$ and $r_g^{(k)} = \sqrt{r_g}$ when $k = 2$. Thus on average, 20% of genes in each dataset were simulated to be differentially expressed.

For each of the 15 simulation settings (3 sample sizes $\times$ 5 variance ratios), 100 independent dataset-pairs were simulated. Gene-wise linear models were fit to each simulated dataset-pair in the statistical computing environment `R` [30] using the model selection methods proposed in subsections 2.1 through 2.3. The accuracy in ranking differentially expressed genes was assessed through the area under the receiver operating characteristic (AUROC) curves using the package `ROCR` [35]. False discovery rate control procedures of Benjamini and Hochberg (BH) [7] and Storey and Tibshirani (ST) [40] were used.

For each method, both results from selecting a variance model for each individual gene and results from selecting an overall variance model for all genes were obtained. For AIC, AICc assuming heterogeneity of variances, BIC, CAIC, HQIC, and the cross-validated prediction log likelihood, the overall model was determined through sums of these criteria. For PRESS, the overall model was determined through gene-wise voting. For our modified MRPP procedure, three p-value cutoffs $1/35 \approx 0.03$, $2/35 \approx 0.06$, or $4/35 \approx 0.11$ were examined, and the overall model was determined by the multivariate test applied to all genes instead of from summarizing gene-wise MRPP tests.

In addition to the variance model selection approaches, the very popular *limma* method [36] was also included as a representative of the variance shrinkage approaches. *Limma* assumes an inverse gamma prior on the gene-wise variances and performs well in our experience. Furthermore, an easy to use `R` package is available for the necessary computation. Although there could be differences in terms of performance between *limma* and other shrinkage methods, we believe the general conclusions would be very similar when compared with our model selection methods. In our simulation study, all default parameter settings in the `limma` package (version 2.18.3) were used.

To further compare our proposed modified MRPP procedure to other commonly used tests of heterogeneity of variances for *univariate* data, we conducted a separate two-sample comparison study, again using subsamples from real microarray data, except that the subsampling of each gene was performed separately to break the correlation across the genes. In this way, the resulting gene-wise p-values across genes can be pooled to provide clearer information about the properties of these tests under the null and under the alternative hypotheses. The alternative tests being compared are Levene's test, the Brown-Forsythe test, the F-test, and a new test, denoted hereafter as reduced Levene's test, which applies ANOVA on the absolute values of the decorrelated and reduced datasets, as in our modified MRPP.

To measure departure from uniformity under the null for each of the tests, the resulting G p-values were first binned into B bins, with bin boundaries determined as the non-redundant set of observed MRPP p-values, together with two natural boundaries 0 and 1. This binning is intended to remove the difference caused by the discrete nature of MRPP p-values, so that results from all tests are comparable. Next, Kullback-Leibler divergence from uniformity as a function of the likelihood ratio was computed for p-values from each method separately, as $\sum_{b=1}^{B} w_b[\log w_b - \log(C_b/G)]$, where w_b is the width of bin b and C_b is the number of p-values falling in bin b. Larger Kullback-Leibler divergence indicates larger departure from uniformity.

Under the alternative hypotheses, samples from one treatment are multiplied by $\sqrt{r_g}$ as before. Because the actual sizes of these approximate tests are not exactly the same, using a fixed cutoff for nominal p-values does not provide fair comparisons. So we computed the probability (*i.e.*, the proportion among the G genes) of the observed test statistic under the alternative hypothesis being at least as extreme as the observed test statistic under the null hypothesis for each simulation. The larger the probability is, the more powerful the test to detect heterogeneities. This measure is in the same spirit as AUROC or the signed rank test. All these simulations were repeated 50 times, and different sample sizes for each treatment were used (Tables 9 to 11).

Our entire simulation process was repeated using three different populations, each derived from a real microarray dataset. The datasets are all publicly available in the GEO database hosted by the National Center for Biotechnology Information (`http://www.ncbi.nlm.nih.gov/geo/`) with accession numbers GSE755 [42], GSE4115 [39] and GSE5406 [18], respectively. The normalized "series" matrix files were downloaded from the GEO website, and the base-2 logarithm was taken to produce our raw data. For each dataset, only the treatment group with the largest sample size was used as the population from which to simulate multiple datasets, as described previously in this subsection.

3. SIMULATION RESULTS AND DISCUSSION

Our simulation results from the three real populations differed only slightly. Thus, we only report results from the population derived from dataset GSE5406 here. To save space, results with different sample sizes are not all reported, but we will mention the trend of change as sample sizes increase.

3.1 Proportion of correctly selected models

The proportion of correctly selected models using each of the model selection procedures is in Tables 2 and 3 for sample sizes 3 and 7, respectively. For all following tables, standard errors are shown in parentheses.

We can see from Table 2 that the PRESS criterion generally prefers larger models in small samples. That is, if the correct model is homogeneous, PRESS has lower probability

Table 2. Proportion ($\times 100$) *of correctly selected models from dataset GSE5406 when sample size* $n_1 = n_2 = 3$

Method	$r_g \equiv 1$		$\mathrm{E}(r_g) = 1$		$\mathrm{E}(r_g) = 3$		$\mathrm{E}(r_g) = 5$		$\mathrm{E}(r_g) = 10$	
	overall	per gene	overall	per gene	overall	per gene	overall	per gene	overall	per gene
AIC	$98.0_{(1.4)}$	$75.7_{(0.4)}$	$4.0_{(2.0)}$	$26.4_{(0.4)}$	$70.0_{(4.6)}$	$41.1_{(1.0)}$	$97.0_{(1.7)}$	$55.3_{(1.1)}$	$100.0_{(0.0)}$	$74.4_{(0.9)}$
AICc	$100.0_{(0.0)}$	$94.7_{(0.2)}$	$0.0_{(0.0)}$	$6.1_{(0.2)}$	$0.0_{(0.0)}$	$13.4_{(0.6)}$	$2.0_{(1.4)}$	$22.9_{(0.9)}$	$39.0_{(4.9)}$	$41.5_{(1.2)}$
BIC	$98.0_{(1.4)}$	$76.6_{(0.4)}$	$1.0_{(1.0)}$	$25.4_{(0.4)}$	$61.0_{(4.9)}$	$40.1_{(1.0)}$	$97.0_{(1.7)}$	$54.3_{(1.1)}$	$100.0_{(0.0)}$	$73.6_{(0.9)}$
CAIC	$100.0_{(0.0)}$	$85.4_{(0.3)}$	$0.0_{(0.0)}$	$16.1_{(0.3)}$	$13.0_{(3.4)}$	$28.9_{(0.9)}$	$53.0_{(5.0)}$	$42.4_{(1.1)}$	$99.0_{(1.0)}$	$63.3_{(1.1)}$
HQIC	$62.0_{(4.8)}$	$68.0_{(0.4)}$	$61.0_{(4.9)}$	$34.1_{(0.4)}$	$100.0_{(0.0)}$	$49.1_{(1.0)}$	$100.0_{(0.0)}$	$62.8_{(1.0)}$	$100.0_{(0.0)}$	$79.9_{(0.8)}$
CV-lik.$_{\text{full}}$	$100.0_{(0.0)}$	$74.1_{(0.3)}$	$0.0_{(0.0)}$	$27.5_{(0.3)}$	$0.0_{(0.0)}$	$38.8_{(0.8)}$	$0.0_{(0.0)}$	$49.8_{(0.9)}$	$13.0_{(3.4)}$	$65.0_{(0.8)}$
PRESS$_{\text{full}}$	$92.0_{(2.7)}$	$55.3_{(0.4)}$	$16.0_{(3.7)}$	$46.6_{(0.4)}$	$85.0_{(3.6)}$	$58.7_{(0.8)}$	$99.0_{(1.0)}$	$69.6_{(0.8)}$	$100.0_{(0.0)}$	$82.9_{(0.6)}$
MRPP$_{0.03}$	$91.0_{(2.9)}$	$95.6_{(0.1)}$	$8.0_{(2.7)}$	$4.9_{(0.1)}$	$70.0_{(4.6)}$	$8.8_{(0.3)}$	$92.0_{(2.7)}$	$13.5_{(0.5)}$	$100.0_{(0.0)}$	$22.6_{(0.6)}$
MRPP$_{0.06}$	$83.0_{(3.8)}$	$91.8_{(0.2)}$	$14.0_{(3.5)}$	$9.0_{(0.2)}$	$81.0_{(3.9)}$	$14.8_{(0.5)}$	$96.0_{(2.0)}$	$21.5_{(0.6)}$	$100.0_{(0.0)}$	$33.3_{(0.7)}$
MRPP$_{0.11}$	$70.0_{(4.6)}$	$84.8_{(0.3)}$	$25.0_{(4.3)}$	$16.4_{(0.3)}$	$88.0_{(3.2)}$	$24.9_{(0.7)}$	$99.0_{(1.0)}$	$34.2_{(0.8)}$	$100.0_{(0.0)}$	$48.9_{(0.8)}$

Table 3. Proportion ($\times 100$) *of correctly selected models from dataset GSE5406 when sample size* $n_1 = n_2 = 7$

Method	$r_g \equiv 1$		$\mathrm{E}(r_g) = 1$		$\mathrm{E}(r_g) = 3$		$\mathrm{E}(r_g) = 5$		$\mathrm{E}(r_g) = 10$	
	overall	per gene	overall	per gene	overall	per gene	overall	per gene	overall	per gene
AIC	$74.0_{(4.4)}$	$72.1_{(0.5)}$	$50.0_{(5.0)}$	$31.9_{(0.4)}$	$100.0_{(0.0)}$	$62.6_{(1.1)}$	$100.0_{(0.0)}$	$82.0_{(0.8)}$	$100.0_{(0.0)}$	$95.6_{(0.3)}$
AICc	$99.0_{(1.0)}$	$78.4_{(0.4)}$	$5.0_{(2.2)}$	$25.4_{(0.4)}$	$98.0_{(1.4)}$	$56.4_{(1.1)}$	$100.0_{(0.0)}$	$77.7_{(0.9)}$	$100.0_{(0.0)}$	$94.1_{(0.4)}$
BIC	$100.0_{(0.0)}$	$82.7_{(0.4)}$	$1.0_{(1.0)}$	$20.9_{(0.4)}$	$91.0_{(2.9)}$	$51.2_{(1.1)}$	$100.0_{(0.0)}$	$73.8_{(1.0)}$	$100.0_{(0.0)}$	$92.6_{(0.5)}$
CAIC	$100.0_{(0.0)}$	$88.0_{(0.3)}$	$0.0_{(0.0)}$	$15.0_{(0.3)}$	$68.0_{(4.7)}$	$43.1_{(1.2)}$	$99.0_{(1.0)}$	$66.9_{(1.1)}$	$100.0_{(0.0)}$	$89.6_{(0.6)}$
HQIC	$96.0_{(2.0)}$	$75.5_{(0.4)}$	$19.0_{(3.9)}$	$28.4_{(0.4)}$	$100.0_{(0.0)}$	$59.4_{(1.1)}$	$100.0_{(0.0)}$	$79.8_{(0.9)}$	$100.0_{(0.0)}$	$94.8_{(0.4)}$
CV-lik.$_{\text{full}}$	$100.0_{(0.0)}$	$75.6_{(0.4)}$	$0.0_{(0.0)}$	$28.2_{(0.4)}$	$81.0_{(3.9)}$	$56.2_{(1.0)}$	$100.0_{(0.0)}$	$75.1_{(0.9)}$	$100.0_{(0.0)}$	$90.5_{(0.5)}$
PRESS$_{\text{full}}$	$96.0_{(2.0)}$	$59.1_{(0.5)}$	$11.0_{(3.1)}$	$45.1_{(0.4)}$	$100.0_{(0.0)}$	$72.5_{(0.9)}$	$100.0_{(0.0)}$	$87.6_{(0.6)}$	$100.0_{(0.0)}$	$97.1_{(0.2)}$
MRPP$_{0.03}$	$93.0_{(2.5)}$	$95.0_{(0.2)}$	$14.0_{(3.5)}$	$6.8_{(0.2)}$	$97.0_{(1.7)}$	$25.2_{(0.8)}$	$100.0_{(0.0)}$	$45.5_{(1.0)}$	$100.0_{(0.0)}$	$73.5_{(0.9)}$
MRPP$_{0.06}$	$87.0_{(3.4)}$	$91.3_{(0.2)}$	$21.0_{(4.1)}$	$11.3_{(0.3)}$	$99.0_{(1.0)}$	$34.8_{(1.0)}$	$100.0_{(0.0)}$	$57.0_{(1.0)}$	$100.0_{(0.0)}$	$82.4_{(0.7)}$
MRPP$_{0.11}$	$77.0_{(4.2)}$	$84.5_{(0.3)}$	$23.0_{(4.2)}$	$19.1_{(0.4)}$	$100.0_{(0.0)}$	$47.1_{(1.0)}$	$100.0_{(0.0)}$	$69.1_{(1.0)}$	$100.0_{(0.0)}$	$89.6_{(0.5)}$

of selecting such a model; but if the correct model is heterogeneous, PRESS performs best to choose the larger model. This is most obvious for gene-by-gene model selection, but when we consider selecting an overall model for all genes by voting, PRESS still picks out the correct model $> 90\%$ of the time even if the true model is homogeneous. Table 3 verifies this and further shows that, when sample size increases, the performance of PRESS also improves accordingly.

On the contrary, the AICc, CAIC and the CV log likelihood criteria seem to prefer smaller models (Table 2). They mostly pick the right model when the true model has homogeneous variances across datasets, but they tend not to choose the correct model when the true variance model is heterogeneous until $\mathrm{E}(r_g)$ is fairly large. However, when sample size increases (Table 3), their performance also improves, in particular AICc. This is reasonable because AICc converges to AIC asymptotically.

The performance of AIC, BIC and HQIC generally lies between the two extremes (Tables 2 and 3). However, when sample size increases, AIC still shows a relatively higher probability of overfitting, whereas the overfitting for HQIC in small samples diminishes as sample size increases. This is because the penalty for AIC is independent of sample size, whereas BIC and HQIC use sample size as auxiliary information to achieve model dimension consistency.

For the model selection criteria other than PRESS and modified MRPP, the use of sum of individual gene-wise criteria seems to have good performance (Tables 2 and 3). For all methods, except CV log likelihood under small sample sizes, the probability of selecting the correct overall model is generally larger than the corresponding probability of selecting the correct individual gene model. This demonstrates the advantage of sharing information across genes, especially when the majority of genes provide concordant information about which model is preferred. The CV log likelihood prefers smaller models too often and using the sum does not help much. However, when sample size increases, it performs similarly to other criteria.

For the modified MRPP test, when the null hypothesis is true, *i.e.*, when the correct model is homogeneous, the probability of selecting the right model is generally slightly smaller than $1 - \alpha$ (Tables 2 and 3), which indicates that the test is only approximate, tending to produce slightly liberal p-values. This problem is more severe for the multivariate version of the test to select an overall model, but less so for univariate tests to select individual gene models. Theoretically, when the responses are indeed normally distributed, the decorrelation in our modified MRPP offers complete independence. The better control for type I error shown in univariate tests than multivariate tests suggests that the marginal distribution of each gene's expression is

probably not far from normality, but the joint distribution of all genes deviates further from multivariate normality due to complicated unknown dependencies among genes. However, when sample size increases, the control of type I error for the multivariate test also gets tighter gradually.

Another phenomenon to note is that the univariate modified MRPP tests do not yield a high proportion of rejected hypotheses when the true model is heterogeneous (Tables 2 and 3), which suggests the use of other model selection criteria for individual gene models. However, when we consider the multivariate test to choose an overall model, the power of the modified MRPP increases considerably. This is because the test statistic we use is very sensitive when the direction of variability change is generally the same for the majority of genes. In this case, the high-dimensionality amplifies such a concordant directional change, and the test power is actually improved by the high dimensionality.

Although there seem to be some shortcomings, *i.e.*, being approximate and not powerful in the univariate case, our modified MRPP test actually suffers less severely from such problems than other tests, as shown in Subsection 3.4.

Lastly, from the results shown in Tables 2 and 3, we see that when the variances across the two datasets are different but $\mathrm{E}(r_g) = 1$, none of the model selection procedures work well. However, this is not necessarily a disadvantage, because even if the correct model is heterogeneous, we need enough data to support the use of such a larger model. If the correct but larger model differs little from the smaller model, the cost in losing precision of estimates may not justify the use of the correct model. This will be further demonstrated in the next subsection.

3.2 Ranking genes in terms of differential expression

Although the proportion of correctly selected models is a good intuitive measure of performance of model selection procedures and is used in many simulation studies, it cannot reveal how well each procedure can pick the differentially expressed genes out of the other tens of thousands of equivalently expressed genes. Hence, we next conduct gene-wise F-tests under the model selected by each procedure to rank the genes according to their p-values and compare the area under the corresponding ROC curves to see how methods differ in terms of detecting differential expression. Results from study GSE5406 when models are selected for each individual gene are shown in Tables 4 and 5, with Table 4 showing results with sample size = 7, and Table 5 showing results with sample size = 3. These results are summarized by contrasts of interest. Contrast k represents the comparison of treatment means within the kth dataset, $k = 1$ or 2. By simulation, the first dataset generally has smaller variance than the second dataset. Further, to ease the direct comparison of the AUROC's across different methods, the standard errors reported in parentheses are computed after removing the common dataset effects by assuming a two-way (dataset & methods) linear model on the AUROC's. Note that although the differences in AUROC values shown in these tables are small, they are still important because 1) compared with standard errors, the difference is usually significant, and 2) the total number of genes is quite large so that even a small difference in AUROC may reflect substantial changes in the rankings of a large number of genes.

Tables 4 and 5 show that using the correct model does not necessarily give the best ranking of genes, even if sample sizes are not small. When $\mathrm{E}(r_g)$ is large, the correct model, *i.e.*, the "separate" row in the tables, generally performs well and the AUROC's are big. But when the true model is heterogeneous but $\mathrm{E}(r_g)$ is close to 1, the correct model is actually the worst in terms of AUROC. In both this case and when the true model is homogeneous, the best procedure is to always perform a pooled analysis of the two datasets, whether it is the correct model or not. A partial explanation is that when $\mathrm{E}(r_g)$ is close to 1, we need much larger sample sizes to get stable estimates from using the larger, but correct, model.

However, this is not the whole story. A special feature shown in Tables 4 and 5 is that, for contrast 2, no matter what the correct model is, performing a pooled analysis is always better than performing a separate analysis, even if $\mathrm{E}(r_g)$ is very large. The same phenomenon is also observed in the other two datasets GSE4115 and GSE755. Further, when we greatly increase the sample size, the same phenomenon still occurs (results not shown). Hence, this is not a phenomenon that can be explained by the lack of stable estimates under small sample sizes.

Since this phenomenon only occurs for the contrast in the second dataset, which has larger variances than the first, one might conjecture that the pooled variance estimate in this case is a shrinkage estimate that actually has smaller squared error risk and/or Stein-type risk. However, this does not fully explain the phenomenon either. A related observation by [43] demonstrated that optimal shrinkage estimation of variances does not always perform best. Our results confirm this observation using real datasets; using pooled estimates of variances is better in terms of AUROC for contrast 2 even if the variance estimates are far from the true parameters in the heterogeneous model. Hence our tentative conclusion is that AUROC is simply a different criterion than the usual loss functions for the variances. The AUROC tries to estimate the probability of giving higher ranking to a random differentially expressed gene than to a random equivalently expressed gene. It may or may not correspond to using good variance estimates, and it does depend upon which contrast is of interest. Further studies on methods optimizing AUROC should be valuable.

One can also see from Tables 4 and 5 that, for the cross-validation procedures, criteria based on full data and criteria only based on specific parts of the data do not necessarily select the same model. For contrast 1, contrast-specific CV

Table 4. Area ($\times 1000$) *under the ROC curve for selecting individual gene model from dataset GSE5406 and* $n_1 = n_2 = 7$

Method	Contrast 1					Contrast 2				
	$r_g \equiv 1$	$E(r_g)=1$	$E(r_g)=3$	$E(r_g)=5$	$E(r_g)=10$	$r_g \equiv 1$	$E(r_g)=1$	$E(r_g)=3$	$E(r_g)=5$	$E(r_g)=10$
AIC	$715.9_{(0.0)}$	$717.6_{(0.0)}$	$716.5_{(0.1)}$	$716.5_{(0.1)}$	$716.9_{(0.0)}$	$717.5_{(0.0)}$	$712.3_{(0.1)}$	$710.8_{(0.0)}$	$710.7_{(0.0)}$	$711.6_{(0.0)}$
AICc	$716.1_{(0.0)}$	$717.6_{(0.0)}$	$716.0_{(0.1)}$	$715.9_{(0.1)}$	$716.7_{(0.0)}$	$717.7_{(0.0)}$	$712.4_{(0.0)}$	$710.6_{(0.0)}$	$710.3_{(0.0)}$	$711.3_{(0.0)}$
BIC	$716.4_{(0.0)}$	$717.7_{(0.0)}$	$715.5_{(0.1)}$	$715.4_{(0.1)}$	$716.4_{(0.0)}$	$717.9_{(0.0)}$	$712.5_{(0.0)}$	$710.4_{(0.0)}$	$709.9_{(0.0)}$	$711.0_{(0.0)}$
CAIC	$716.7_{(0.0)}$	$717.8_{(0.0)}$	$714.8_{(0.1)}$	$714.2_{(0.1)}$	$715.5_{(0.1)}$	$718.1_{(0.0)}$	$712.5_{(0.0)}$	$710.3_{(0.1)}$	$709.3_{(0.1)}$	$710.4_{(0.0)}$
HQIC	$716.0_{(0.0)}$	$717.6_{(0.0)}$	$716.3_{(0.1)}$	$716.2_{(0.1)}$	$716.8_{(0.0)}$	$717.6_{(0.0)}$	$712.4_{(0.0)}$	$710.7_{(0.0)}$	$710.5_{(0.0)}$	$711.4_{(0.0)}$
CV-lik.$_{\text{full}}$	$716.0_{(0.0)}$	$717.6_{(0.0)}$	$715.3_{(0.1)}$	$714.5_{(0.1)}$	$714.5_{(0.1)}$	$717.6_{(0.0)}$	$712.2_{(0.0)}$	$710.7_{(0.0)}$	$710.2_{(0.0)}$	$710.8_{(0.0)}$
CV-lik.$_{\text{specific}}$	$715.9_{(0.0)}$	$717.5_{(0.0)}$	$714.6_{(0.1)}$	$713.1_{(0.1)}$	$712.7_{(0.1)}$	$717.5_{(0.0)}$	$712.2_{(0.0)}$	$711.2_{(0.0)}$	$711.2_{(0.0)}$	$711.8_{(0.0)}$
PRESS$_{\text{full}}$	$715.5_{(0.0)}$	$717.4_{(0.1)}$	$716.9_{(0.1)}$	$716.8_{(0.0)}$	$717.1_{(0.0)}$	$717.2_{(0.1)}$	$712.2_{(0.1)}$	$711.3_{(0.0)}$	$711.4_{(0.0)}$	$711.9_{(0.0)}$
PRESS$_{\text{specific}}$	$716.0_{(0.1)}$	$717.7_{(0.1)}$	$717.5_{(0.1)}$	$717.3_{(0.0)}$	$717.2_{(0.0)}$	$717.7_{(0.1)}$	$712.6_{(0.1)}$	$710.9_{(0.1)}$	$710.5_{(0.1)}$	$711.1_{(0.1)}$
MRPP$_{0.03}$	$717.3_{(0.1)}$	$718.2_{(0.1)}$	$712.6_{(0.2)}$	$708.6_{(0.2)}$	$707.3_{(0.2)}$	$718.4_{(0.1)}$	$712.6_{(0.1)}$	$711.1_{(0.1)}$	$709.0_{(0.1)}$	$708.2_{(0.1)}$
MRPP$_{0.06}$	$717.0_{(0.1)}$	$718.0_{(0.1)}$	$713.2_{(0.1)}$	$710.9_{(0.1)}$	$711.3_{(0.1)}$	$718.1_{(0.1)}$	$712.6_{(0.1)}$	$710.7_{(0.1)}$	$709.2_{(0.1)}$	$709.3_{(0.1)}$
MRPP$_{0.11}$	$716.5_{(0.1)}$	$717.7_{(0.0)}$	$714.4_{(0.1)}$	$713.5_{(0.1)}$	$714.6_{(0.1)}$	$717.9_{(0.1)}$	$712.5_{(0.1)}$	$710.7_{(0.0)}$	$709.8_{(0.0)}$	$710.5_{(0.0)}$
separate	$715.2_{(0.1)}$	$717.3_{(0.1)}$	$717.3_{(0.1)}$	$717.3_{(0.0)}$	$717.3_{(0.0)}$	$716.9_{(0.1)}$	$712.2_{(0.1)}$	$712.2_{(0.0)}$	$712.2_{(0.0)}$	$712.2_{(0.0)}$
pool	$717.9_{(0.1)}$	$718.4_{(0.1)}$	$715.4_{(0.2)}$	$713.7_{(0.3)}$	$711.8_{(0.3)}$	$718.9_{(0.1)}$	$712.8_{(0.2)}$	$714.1_{(0.1)}$	$714.0_{(0.1)}$	$713.5_{(0.1)}$
limma	$715.7_{(0.1)}$	$717.8_{(0.1)}$	$717.8_{(0.1)}$	$717.8_{(0.1)}$	$717.8_{(0.1)}$	$718.5_{(0.1)}$	$713.3_{(0.1)}$	$713.3_{(0.1)}$	$713.3_{(0.1)}$	$713.3_{(0.1)}$

Table 5. Area ($\times 1000$) *under the ROC curve for selecting individual gene model from dataset GSE5406 and* $n_1 = n_2 = 3$

Method	Contrast 1					Contrast 2				
	$r_g \equiv 1$	$E(r_g)=1$	$E(r_g)=3$	$E(r_g)=5$	$E(r_g)=10$	$r_g \equiv 1$	$E(r_g)=1$	$E(r_g)=3$	$E(r_g)=5$	$E(r_g)=10$
AIC	$641.0_{(0.1)}$	$631.2_{(0.1)}$	$636.3_{(0.1)}$	$635.2_{(0.2)}$	$634.9_{(0.2)}$	$636.1_{(0.1)}$	$637.3_{(0.1)}$	$633.9_{(0.1)}$	$631.8_{(0.1)}$	$630.7_{(0.1)}$
AICc	$643.0_{(0.1)}$	$632.9_{(0.1)}$	$634.3_{(0.2)}$	$630.1_{(0.2)}$	$625.4_{(0.2)}$	$638.3_{(0.1)}$	$638.7_{(0.1)}$	$637.2_{(0.1)}$	$634.2_{(0.1)}$	$629.5_{(0.2)}$
BIC	$641.0_{(0.1)}$	$631.2_{(0.1)}$	$636.2_{(0.1)}$	$635.1_{(0.2)}$	$634.7_{(0.2)}$	$636.2_{(0.1)}$	$637.4_{(0.1)}$	$633.9_{(0.1)}$	$631.7_{(0.1)}$	$630.6_{(0.1)}$
CAIC	$641.6_{(0.1)}$	$631.8_{(0.1)}$	$635.2_{(0.1)}$	$633.0_{(0.1)}$	$631.9_{(0.2)}$	$636.9_{(0.1)}$	$637.9_{(0.1)}$	$634.9_{(0.1)}$	$632.0_{(0.1)}$	$629.5_{(0.1)}$
HQIC	$640.6_{(0.1)}$	$630.8_{(0.1)}$	$637.0_{(0.1)}$	$636.3_{(0.2)}$	$635.9_{(0.2)}$	$635.6_{(0.1)}$	$637.1_{(0.1)}$	$633.6_{(0.1)}$	$632.0_{(0.1)}$	$631.5_{(0.1)}$
CV-lik.$_{\text{full}}$	$641.2_{(0.1)}$	$631.5_{(0.1)}$	$635.2_{(0.1)}$	$632.6_{(0.1)}$	$629.1_{(0.2)}$	$636.4_{(0.1)}$	$637.4_{(0.1)}$	$634.6_{(0.1)}$	$632.5_{(0.1)}$	$630.7_{(0.1)}$
CV-lik.$_{\text{specific}}$	$640.9_{(0.1)}$	$631.1_{(0.1)}$	$634.7_{(0.1)}$	$631.5_{(0.2)}$	$626.8_{(0.2)}$	$636.0_{(0.1)}$	$637.2_{(0.1)}$	$634.3_{(0.1)}$	$632.9_{(0.1)}$	$632.2_{(0.1)}$
PRESS$_{\text{full}}$	$640.2_{(0.1)}$	$630.5_{(0.1)}$	$636.8_{(0.1)}$	$636.0_{(0.2)}$	$635.3_{(0.2)}$	$635.3_{(0.1)}$	$636.8_{(0.1)}$	$633.7_{(0.1)}$	$632.6_{(0.1)}$	$632.2_{(0.1)}$
PRESS$_{\text{specific}}$	$640.8_{(0.1)}$	$631.0_{(0.1)}$	$638.1_{(0.2)}$	$637.6_{(0.2)}$	$637.2_{(0.2)}$	$635.9_{(0.1)}$	$637.4_{(0.1)}$	$634.4_{(0.1)}$	$632.8_{(0.1)}$	$631.2_{(0.1)}$
MRPP$_{0.03}$	$643.6_{(0.1)}$	$633.4_{(0.1)}$	$635.7_{(0.2)}$	$631.1_{(0.3)}$	$622.9_{(0.5)}$	$638.8_{(0.1)}$	$639.3_{(0.1)}$	$638.5_{(0.1)}$	$636.6_{(0.1)}$	$633.2_{(0.1)}$
MRPP$_{0.06}$	$643.0_{(0.1)}$	$633.0_{(0.1)}$	$635.0_{(0.2)}$	$630.2_{(0.3)}$	$622.6_{(0.5)}$	$638.2_{(0.1)}$	$638.9_{(0.1)}$	$637.5_{(0.1)}$	$635.2_{(0.1)}$	$631.8_{(0.1)}$
MRPP$_{0.11}$	$642.2_{(0.1)}$	$632.3_{(0.1)}$	$634.5_{(0.2)}$	$630.5_{(0.3)}$	$624.8_{(0.3)}$	$637.5_{(0.1)}$	$638.3_{(0.1)}$	$636.0_{(0.1)}$	$633.5_{(0.1)}$	$630.5_{(0.1)}$
separate	$639.1_{(0.2)}$	$629.6_{(0.2)}$	$637.0_{(0.2)}$	$637.0_{(0.2)}$	$637.0_{(0.2)}$	$634.3_{(0.2)}$	$635.9_{(0.1)}$	$634.9_{(0.1)}$	$634.9_{(0.1)}$	$634.9_{(0.1)}$
pool	$644.4_{(0.2)}$	$634.0_{(0.2)}$	$638.1_{(0.3)}$	$636.1_{(0.3)}$	$633.5_{(0.4)}$	$639.7_{(0.2)}$	$640.1_{(0.2)}$	$640.5_{(0.1)}$	$639.8_{(0.1)}$	$638.6_{(0.1)}$
limma	$643.1_{(0.3)}$	$633.2_{(0.3)}$	$641.1_{(0.4)}$	$641.1_{(0.4)}$	$641.1_{(0.4)}$	$641.0_{(0.2)}$	$642.1_{(0.3)}$	$641.1_{(0.2)}$	$641.1_{(0.2)}$	$641.1_{(0.3)}$

log likelihood deteriorates the AUROC compared with the full CV log likelihood, whereas contrast-specific PRESS improves AUROC compared with the full PRESS. However, for contrast 2, this is further complicated by $E(r_g)$; when $E(r_g)$ is large, the contrast-specific CV log likelihood also improves AUROC, but on the contrary, contrast-specific PRESS produces worse results this time; and the conclusion reverses when $E(r_g) = 1$ or is close to 1. Hence, these cross-validation procedures that are based on only part of the data are preferred to the ordinary ones only under some but not all situations, and care has to be taken in practice to choose a good procedure.

Conclusions from selecting an overall model for all genes (data not shown) generally agree with those from gene-wise model selection (Table 4). Moreover, when $E(r_g)$ is larger, most of the model selection procedures will choose the correct model, and hence their differences in AUROC's are largely indiscernible from each other.

When comparing model selection procedures with the variance shrinkage procedure *limma*, sample size plays an important role. When sample size is small (Table 5), variance shrinkage often outperforms model selection procedures by a large margin, except in a few cases where it is slightly worse than the best performing procedure. On the other hand, when sample sizes increase (Table 4), the advantage of variance shrinkage gradually fades away and shrinkage performs similarly to model selection procedures. This phenomenon is intuitive, because when sample sizes are small, the total information contained in the additional data is still relatively scarce compared with the information contained in the large number of genes from a single dataset. In this case, pooling information across genes is

Table 6. False discovery proportions and number of rejections for contrast 1 from individual gene model selection on dataset GSE5406 using the BH method when $n_1 = n_2 = 3$ and $r \equiv 1$ for all genes

Method	FDR = 5%		FDR = 10%		FDR = 15%		FDR = 20%	
	FDP%	#Rej	FDP%	#Rej	FDP%	#Rej	FDP%	#Rej
AIC	$5.9_{(1.1)}$	$10.8_{(2.4)}$	$11.6_{(1.4)}$	$56.2_{(14.1)}$	$16.9_{(1.6)}$	$143.3_{(31.0)}$	$21.2_{(1.6)}$	$264.1_{(50.0)}$
AICc	$6.7_{(1.2)}$	$13.3_{(2.4)}$	$11.5_{(1.4)}$	$53.6_{(10.4)}$	$16.7_{(1.5)}$	$133.8_{(24.7)}$	$20.4_{(1.5)}$	$242.6_{(43.0)}$
BIC	$5.9_{(1.1)}$	$11.2_{(2.4)}$	$11.6_{(1.4)}$	$56.8_{(14.2)}$	$16.7_{(1.6)}$	$144.2_{(30.9)}$	$21.2_{(1.6)}$	$265.7_{(50.0)}$
CAIC	$6.6_{(1.2)}$	$13.1_{(2.7)}$	$11.7_{(1.4)}$	$60.1_{(13.6)}$	$17.3_{(1.6)}$	$150.5_{(29.6)}$	$21.1_{(1.6)}$	$271.3_{(48.4)}$
HQIC	$5.7_{(1.2)}$	$9.3_{(1.9)}$	$11.1_{(1.4)}$	$50.2_{(13.6)}$	$16.6_{(1.6)}$	$131.6_{(30.4)}$	$21.0_{(1.7)}$	$249.8_{(50.0)}$
CV-lik.$_{\text{full}}$	$5.5_{(1.0)}$	$9.6_{(1.7)}$	$11.3_{(1.6)}$	$45.3_{(10.5)}$	$15.7_{(1.6)}$	$119.0_{(26.3)}$	$20.1_{(1.6)}$	$226.3_{(44.6)}$
CV-lik.$_{\text{specific}}$	$5.4_{(1.4)}$	$6.9_{(1.2)}$	$10.6_{(1.6)}$	$32.3_{(\ 7.8)}$	$15.2_{(1.7)}$	$88.5_{(20.9)}$	$19.1_{(1.6)}$	$168.4_{(35.3)}$
PRESS$_{\text{full}}$	$5.6_{(1.4)}$	$7.0_{(1.3)}$	$11.0_{(1.7)}$	$35.9_{(\ 9.8)}$	$16.9_{(1.9)}$	$100.2_{(26.0)}$	$21.0_{(1.8)}$	$198.8_{(44.3)}$
PRESS$_{\text{specific}}$	$4.7_{(1.0)}$	$5.9_{(0.9)}$	$12.7_{(1.7)}$	$38.2_{(10.5)}$	$16.7_{(1.8)}$	$110.1_{(28.8)}$	$21.4_{(1.9)}$	$233.2_{(52.4)}$
MRPP$_{0.03}$	$5.7_{(1.0)}$	$9.9_{(1.6)}$	$9.3_{(1.3)}$	$42.1_{(\ 8.1)}$	$13.6_{(1.4)}$	$109.8_{(21.6)}$	$17.6_{(1.5)}$	$205.9_{(39.5)}$
MRPP$_{0.06}$	$6.0_{(1.1)}$	$10.2_{(1.7)}$	$10.2_{(1.3)}$	$44.4_{(\ 9.0)}$	$14.7_{(1.5)}$	$116.0_{(23.4)}$	$18.4_{(1.6)}$	$216.6_{(40.9)}$
MRPP$_{0.11}$	$6.0_{(1.1)}$	$10.3_{(1.8)}$	$11.4_{(1.4)}$	$47.9_{(10.3)}$	$15.2_{(1.5)}$	$124.2_{(25.6)}$	$19.2_{(1.6)}$	$228.3_{(43.1)}$
separate	$3.8_{(1.6)}$	$0.6_{(0.1)}$	$4.8_{(1.5)}$	$3.4_{(\ 1.8)}$	$11.9_{(2.4)}$	$14.0_{(\ 9.2)}$	$13.4_{(2.3)}$	$34.9_{(20.3)}$
pool	$3.4_{(0.7)}$	$8.4_{(1.4)}$	$6.0_{(1.0)}$	$31.0_{(\ 6.0)}$	$9.4_{(1.4)}$	$87.0_{(17.7)}$	$12.4_{(1.5)}$	$174.7_{(36.4)}$
limma	$4.0_{(1.5)}$	$14.6_{(5.4)}$	$5.4_{(1.4)}$	$54.2_{(17.1)}$	$8.3_{(1.6)}$	$119.1_{(34.4)}$	$10.0_{(1.7)}$	$200.7_{(53.6)}$

more beneficial. On other hand, when sample sizes increase (or when more datasets are incorporated into the joint analysis), more and more information can be obtained from additional datasets and the estimates of gene-wise variance become more and more precise. Of course, the number of genes within the dataset from a single study does not increase for a fixed microarray platform, so there is no more information to be borrowed from other genes within a single dataset. Hence shrinkage is not necessarily preferred in such cases. Although we cannot arbitrarily increase sample size in our simulations due the constraints of total population sample size, it is reasonable to believe that when we greatly increase the sample size by including more datasets for a joint analysis, shrinkage methods would eventually have little impact on the results.

So, our general recommendation for pooling variance estimates across datasets to rank genes from most significant to least significant is 1) to use pooled analysis when variances are judged using our methods to be approximately equal across datasets, 2) to use a pooled analysis even when variances differ across datasets if the contrasts of interest are within the datasets with larger variance, and 3) to use a separate analysis when variances differ across datasets and the contrasts of interest are within the datasets with smaller variances. In our experience, shrinking variance estimates by borrowing information across genes is seldom harmful and usually helpful. Thus, whether gene-specific variance estimates are obtained by pooling across datasets or not, we recommend shrinking gene-specific variance estimates by borrowing information across genes using a procedure like *limma*.

3.3 Control of FDR

For the massive amount of hypothesis testing in microarray data, controlling family-wise type I errors is too conservative. One often chooses to control FDR at some prespecified level. It is of interest to see whether FDR is still under control after selecting the variance part of the model. Selected results of the mean false discovery proportions (FDP%) and the number of rejected hypotheses (#Rej) are shown in Tables 6 to 8.

Table 6 shows results when the true model is homogeneous and sample size is small for the contrast in dataset 1, using the BH procedure and with model selection performed on a gene-by-gene basis. We see that always using the correct model, *i.e.*, the pooled analysis, generally controls the FDR below the desired level. Although always using the separate analysis also successfully controls the FDR, the number of rejections (*i.e.*, power) is much lower compared with using the correct model.

Also note in Table 6 that, although both the separate analysis and the pooled analysis control FDR, this is no longer the case when some genes use a separate analysis and others use the pooled analysis. For most model selection procedures, gene-by-gene model selection tends to increase the false discovery proportions. Because the BH procedure does not estimate the proportion of null hypotheses, it is generally considered as a conservative method compared to methods that are more adaptive, *e.g.*, the ST procedure. Not surprisingly, if we use the ST procedure after gene-by-gene model selection, the actual FDPs increase further above the desired level (results not shown).

However, as sample size increases (Table 7), the liberalness of p-values is largely alleviated and the control of FDR is still successful for BH, even if after gene-wise model selection. The ST method is still slightly liberal when sample size increases, but the severity is lower and probably ignorable in practice (results not shown). Moreover, variance shrinkage through *limma* generally outperforms model selection when sample size is small (Table 6), but is not as powerful

Table 7. False discovery proportions and number of rejections for contrast 1 from individual gene model selection on dataset GSE5406 using the BH method when $n_1 = n_2 = 7$ and $r \equiv 1$ for all genes

Method	FDR = 5%		FDR = 10%		FDR = 15%		FDR = 20%	
	FDP%	#Rej	FDP%	#Rej	FDP%	#Rej	FDP%	#Rej
AIC	$4.8_{(0.5)}$	$479.8_{(17.2)}$	$9.1_{(0.9)}$	$737.5_{(27.0)}$	$13.2_{(1.1)}$	$977.2_{(39.1)}$	$17.1_{(1.3)}$	$1220.1_{(53.1)}$
AICc	$4.8_{(0.5)}$	$485.3_{(16.5)}$	$9.0_{(0.8)}$	$740.5_{(26.1)}$	$13.1_{(1.1)}$	$980.4_{(37.7)}$	$17.1_{(1.2)}$	$1225.0_{(51.9)}$
BIC	$4.8_{(0.5)}$	$487.5_{(16.1)}$	$8.9_{(0.8)}$	$740.0_{(25.4)}$	$13.0_{(1.1)}$	$978.6_{(36.9)}$	$16.9_{(1.2)}$	$1224.2_{(51.4)}$
CAIC	$4.6_{(0.5)}$	$485.8_{(15.6)}$	$8.7_{(0.8)}$	$735.6_{(24.6)}$	$12.8_{(1.1)}$	$970.3_{(35.8)}$	$16.6_{(1.2)}$	$1214.1_{(50.4)}$
HQIC	$4.8_{(0.6)}$	$483.4_{(16.9)}$	$9.1_{(0.9)}$	$740.8_{(26.6)}$	$13.2_{(1.1)}$	$979.8_{(38.3)}$	$17.1_{(1.3)}$	$1224.5_{(52.5)}$
CV-lik.$_{\text{full}}$	$4.7_{(0.6)}$	$476.4_{(16.6)}$	$8.9_{(0.9)}$	$728.1_{(26.3)}$	$13.0_{(1.1)}$	$966.7_{(38.2)}$	$17.0_{(1.3)}$	$1212.8_{(52.7)}$
CV-lik.$_{\text{specific}}$	$4.3_{(0.5)}$	$439.4_{(16.1)}$	$8.3_{(0.8)}$	$677.9_{(25.0)}$	$12.0_{(1.0)}$	$896.6_{(35.3)}$	$15.8_{(1.2)}$	$1125.5_{(48.4)}$
PRESS$_{\text{full}}$	$4.6_{(0.5)}$	$453.3_{(17.6)}$	$8.9_{(0.8)}$	$707.2_{(27.6)}$	$13.0_{(1.1)}$	$946.7_{(40.1)}$	$16.8_{(1.3)}$	$1189.7_{(54.7)}$
PRESS$_{\text{specific}}$	$5.0_{(0.6)}$	$497.2_{(17.9)}$	$9.6_{(0.9)}$	$778.1_{(29.4)}$	$14.1_{(1.1)}$	$1044.9_{(44.0)}$	$18.4_{(1.3)}$	$1320.2_{(60.8)}$
MRPP$_{0.03}$	$4.2_{(0.5)}$	$465.2_{(15.0)}$	$8.0_{(0.8)}$	$707.8_{(24.1)}$	$11.9_{(1.1)}$	$933.6_{(35.5)}$	$15.6_{(1.3)}$	$1173.2_{(50.7)}$
MRPP$_{0.06}$	$4.4_{(0.5)}$	$470.7_{(15.3)}$	$8.4_{(0.8)}$	$717.2_{(24.8)}$	$12.3_{(1.1)}$	$947.3_{(36.2)}$	$16.1_{(1.3)}$	$1189.5_{(51.2)}$
MRPP$_{0.11}$	$4.6_{(0.5)}$	$474.3_{(15.8)}$	$8.7_{(0.8)}$	$724.3_{(25.6)}$	$12.7_{(1.1)}$	$957.7_{(37.5)}$	$16.5_{(1.3)}$	$1202.8_{(52.4)}$
separate	$3.5_{(0.5)}$	$296.3_{(16.4)}$	$6.6_{(0.7)}$	$522.9_{(25.6)}$	$10.2_{(0.9)}$	$736.5_{(36.6)}$	$13.6_{(1.1)}$	$954.9_{(49.8)}$
pool	$3.5_{(0.5)}$	$442.5_{(14.8)}$	$7.1_{(0.8)}$	$676.8_{(23.5)}$	$10.8_{(1.1)}$	$897.6_{(35.0)}$	$14.5_{(1.3)}$	$1130.7_{(50.1)}$
limma	$3.2_{(0.5)}$	$371.4_{(19.4)}$	$6.4_{(0.8)}$	$601.1_{(29.8)}$	$10.1_{(1.1)}$	$823.9_{(41.7)}$	$13.6_{(1.3)}$	$1047.4_{(56.3)}$

Table 8. False discovery proportions and number of rejections for contrast 2 from individual gene model selection on dataset GSE5406 using the BH method when $n_1 = n_2 = 3$ and $\mathrm{E}(r_g) = 10$ for all genes

Method	FDR = 5%		FDR = 10%		FDR = 15%		FDR = 20%	
	FDP%	#Rej	FDP%	#Rej	FDP%	#Rej	FDP%	#Rej
AIC	$31.9_{(1.3)}$	$164.6_{(13.7)}$	$39.2_{(1.2)}$	$332.1_{(26.1)}$	$43.5_{(1.2)}$	$498.7_{(38.0)}$	$46.8_{(1.1)}$	$674.7_{(51.1)}$
AICc	$32.8_{(1.3)}$	$386.4_{(29.2)}$	$41.6_{(1.2)}$	$795.8_{(54.7)}$	$47.1_{(1.2)}$	$1204.4_{(77.7)}$	$51.0_{(1.1)}$	$1632.3_{(100.6)}$
BIC	$31.9_{(1.3)}$	$171.2_{(14.1)}$	$39.4_{(1.2)}$	$343.8_{(26.8)}$	$43.7_{(1.2)}$	$515.6_{(39.0)}$	$47.0_{(1.1)}$	$698.0_{(52.5)}$
CAIC	$32.3_{(1.3)}$	$244.5_{(19.3)}$	$40.1_{(1.2)}$	$494.6_{(36.1)}$	$45.2_{(1.1)}$	$742.8_{(52.4)}$	$48.6_{(1.1)}$	$996.3_{(69.0)}$
HQIC	$31.1_{(1.4)}$	$122.6_{(10.6)}$	$38.7_{(1.2)}$	$250.4_{(20.6)}$	$42.6_{(1.1)}$	$383.1_{(30.7)}$	$45.5_{(1.1)}$	$523.8_{(42.2)}$
CV-lik.$_{\text{full}}$	$31.2_{(1.3)}$	$182.9_{(15.2)}$	$38.5_{(1.2)}$	$388.8_{(30.0)}$	$43.3_{(1.2)}$	$603.7_{(45.0)}$	$46.5_{(1.1)}$	$831.8_{(60.9)}$
CV-lik.$_{\text{specific}}$	$29.4_{(1.4)}$	$90.8_{(\ 8.4)}$	$36.7_{(1.2)}$	$193.8_{(17.1)}$	$40.5_{(1.1)}$	$299.9_{(25.9)}$	$43.5_{(1.1)}$	$423.1_{(36.3)}$
PRESS$_{\text{full}}$	$28.9_{(1.5)}$	$83.7_{(\ 7.9)}$	$36.6_{(1.2)}$	$181.6_{(16.1)}$	$40.3_{(1.2)}$	$286.9_{(25.1)}$	$43.2_{(1.1)}$	$406.7_{(35.6)}$
PRESS$_{\text{specific}}$	$29.5_{(1.3)}$	$172.3_{(14.7)}$	$37.3_{(1.2)}$	$394.9_{(31.5)}$	$42.1_{(1.2)}$	$639.1_{(48.8)}$	$45.9_{(1.1)}$	$910.8_{(68.3)}$
MRPP$_{0.03}$	$32.5_{(1.3)}$	$441.8_{(33.3)}$	$41.3_{(1.2)}$	$959.6_{(65.9)}$	$47.0_{(1.2)}$	$1487.6_{(94.5)}$	$51.4_{(1.1)}$	$2048.0_{(121.9)}$
MRPP$_{0.06}$	$32.5_{(1.3)}$	$376.0_{(28.8)}$	$40.9_{(1.2)}$	$809.9_{(55.8)}$	$46.2_{(1.2)}$	$1250.3_{(81.0)}$	$50.3_{(1.1)}$	$1723.9_{(106.2)}$
MRPP$_{0.11}$	$31.7_{(1.3)}$	$282.6_{(22.1)}$	$39.7_{(1.2)}$	$598.2_{(43.3)}$	$44.9_{(1.2)}$	$927.5_{(63.7)}$	$48.7_{(1.1)}$	$1279.2_{(85.1)}$
separate	$5.0_{(1.9)}$	$0.3_{(\ 0.1)}$	$8.7_{(2.5)}$	$0.6_{(\ 0.1)}$	$10.0_{(2.4)}$	$1.6_{(\ 0.4)}$	$12.2_{(2.2)}$	$5.7_{(\ 1.5)}$
pool	$32.6_{(1.3)}$	$557.9_{(42.1)}$	$42.0_{(1.3)}$	$1239.8_{(84.5)}$	$48.2_{(1.2)}$	$1958.4_{(121.0)}$	$52.9_{(1.2)}$	$2717.2_{(153.9)}$
limma	$1.9_{(1.2)}$	$2.6_{(\ 0.7)}$	$5.0_{(1.5)}$	$19.6_{(\ 5.3)}$	$6.9_{(1.4)}$	$50.6_{(12.7)}$	$8.7_{(1.5)}$	$103.2_{(23.8)}$

as the best model selection procedures when sample size increases (Table 7). This is consistent with the observation on ranking genes in the previous subsection.

When $\mathrm{E}(r_g)$ increases to 10, the FDRs are all controlled at the desired levels, even after gene-by-gene model selection (results not shown). However, the pooled analysis is extremely under powered for contrasts in dataset 1. This is because dataset 1 is simulated with smaller variances than dataset 2 and the pooled analysis for contrasts in dataset 1 always vastly overestimates the variances and reduces power. At higher FDR levels, the CV log likelihood and the MRPP procedure for gene-by-gene analysis result in lower power than the other model selection methods, probably due to their preference for the smaller model under this simulation situation. The variance shrinkage approach is not seriously affected by $\mathrm{E}(r_g)$ but is affected mainly by sample size.

However, when we consider the contrast within the second dataset, where the true variances are much larger than the first, the only method that controls the FDR properly uses a separate analysis for all genes (Table 8). Pooled analysis and all model selection procedures produce mean FDP far above the desired level. Note, however, that in this situation the pooled analysis is able to produce the best ranking of genes, in terms of AUROC.

When variance model selection is conducted on the whole genome scale, the final FDP is largely dependent on how often the procedure chooses the separate model. For example, in the case shown in Table 8 but with model selection methods applied to all genes together, only AICc and CV log

Table 9. Estimated Kullback-Leibler divergence of nominal to theoretical p-value distributions

Sample sizes	Kullback-Leibler divergence ×1000				
	Lev	RLev	BF	F	MRPP
3,3	$88.0_{(0.4)}$	$3.0_{(0.1)}$	$104.4_{(0.5)}$	$4.4_{(0.1)}$	$2.0_{(0.1)}$
3,4	$67.2_{(0.3)}$	$4.7_{(0.1)}$	$18.2_{(0.2)}$	$7.7_{(0.1)}$	$2.9_{(0.1)}$
3,7	$40.9_{(0.3)}$	$6.0_{(0.1)}$	$106.8_{(0.8)}$	$10.6_{(0.1)}$	$2.9_{(0.1)}$
3,10	$32.4_{(0.2)}$	$8.2_{(0.1)}$	$34.4_{(0.3)}$	$11.7_{(0.1)}$	$3.0_{(0.1)}$
4,4	$48.2_{(0.3)}$	$4.9_{(0.1)}$	$2.9_{(0.1)}$	$11.0_{(0.1)}$	$3.7_{(0.1)}$
4,7	$31.8_{(0.2)}$	$7.6_{(0.1)}$	$11.8_{(0.1)}$	$18.8_{(0.2)}$	$5.8_{(0.1)}$
4,10	$28.9_{(0.2)}$	$11.4_{(0.1)}$	$9.7_{(0.1)}$	$24.2_{(0.2)}$	$8.3_{(0.1)}$
7,7	$31.4_{(0.2)}$	$18.1_{(0.1)}$	$46.3_{(0.5)}$	$41.3_{(0.3)}$	$16.6_{(0.1)}$
7,10	$27.1_{(0.2)}$	$17.9_{(0.2)}$	$22.2_{(0.2)}$	$46.5_{(0.2)}$	$17.3_{(0.1)}$

Table 10. Probability that the test statistic under the alternative is at least as extreme as the test statistic under the null when $\mathrm{E}(r_g) = 1$

Sample sizes	Probability ×1000				
	Lev	RLev	BF	F	MRPP
3,3	$530.2_{(0.4)}$	$530.0_{(0.4)}$	$530.4_{(0.5)}$	$530.4_{(0.5)}$	$530.0_{(0.4)}$
3,4	$527.7_{(0.4)}$	$531.1_{(0.4)}$	$520.6_{(0.4)}$	$533.3_{(0.4)}$	$531.0_{(0.4)}$
3,7	$523.6_{(0.5)}$	$532.5_{(0.5)}$	$515.7_{(0.5)}$	$534.8_{(0.5)}$	$532.2_{(0.5)}$
3,10	$521.3_{(0.5)}$	$532.5_{(0.5)}$	$510.8_{(0.5)}$	$535.7_{(0.4)}$	$532.2_{(0.5)}$
4,4	$538.0_{(0.5)}$	$537.6_{(0.5)}$	$537.5_{(0.4)}$	$538.4_{(0.5)}$	$537.6_{(0.5)}$
4,7	$535.4_{(0.5)}$	$540.9_{(0.5)}$	$535.8_{(0.5)}$	$541.9_{(0.5)}$	$540.6_{(0.5)}$
4,10	$533.7_{(0.5)}$	$540.7_{(0.4)}$	$531.3_{(0.4)}$	$542.1_{(0.4)}$	$540.4_{(0.4)}$
7,7	$552.8_{(0.5)}$	$552.1_{(0.6)}$	$553.1_{(0.5)}$	$552.8_{(0.5)}$	$552.2_{(0.6)}$
7,10	$553.5_{(0.4)}$	$555.3_{(0.5)}$	$550.9_{(0.4)}$	$556.1_{(0.4)}$	$555.1_{(0.5)}$

likelihood choose the pooled model frequently, and hence their AUROC's are better but FDR is not controlled. Other methods mostly choose a separate model, and their FDR is below the desired level, although their lists of differentially expressed genes are worse than using the pooled model.

Hence, one faces the dilemma of whether we should care more about gene ranking or care more about control of FDR. Similar to the recommendation in [12], we suggest the following strategy: when combined sample size is small or when model selection procedures suggest that the variances are too different to be combined across studies, use a separate analysis to determine the number of genes that can be declared differentially expressed and a pooled analysis to determine which of the genes are declared as differentially expressed. This hybrid approach will both control FDR well below the desired level and provide a good list of candidate genes for further study.

3.4 Size and power of modified MRPP compared with alternative univariate tests

Among other model selection procedures considered in this study, our modified MRPP is the only one that is based on hypothesis testing, and behaves largely different than other information criteria based methods. Hence it is more reasonable to compare our modified MRPP with other commonly used hypothesis tests for unequal variances in the univariate setting, where all such tests are applicable.

Table 11. Probability that the test statistic under the alternative is at least as extreme as the test statistic under the null when $\mathrm{E}(r_g) = 10$

Sample sizes	Probability ×1000				
	Lev	RLev	BF	F	MRPP
3,3	$730.7_{(0.4)}$	$724.6_{(0.5)}$	$732.0_{(0.4)}$	$733.0_{(0.5)}$	$730.1_{(0.5)}$
3,4	$775.1_{(0.3)}$	$743.1_{(0.4)}$	$826.4_{(0.3)}$	$754.4_{(0.4)}$	$768.0_{(0.4)}$
3,7	$827.8_{(0.4)}$	$769.8_{(0.4)}$	$876.1_{(0.3)}$	$782.6_{(0.4)}$	$816.1_{(0.4)}$
3,10	$850.3_{(0.3)}$	$782.8_{(0.4)}$	$904.4_{(0.2)}$	$796.9_{(0.4)}$	$836.5_{(0.3)}$
4,4	$773.8_{(0.5)}$	$768.2_{(0.5)}$	$769.6_{(0.5)}$	$778.8_{(0.4)}$	$774.9_{(0.5)}$
4,7	$832.4_{(0.4)}$	$798.0_{(0.4)}$	$829.0_{(0.4)}$	$815.1_{(0.3)}$	$831.8_{(0.4)}$
4,10	$854.4_{(0.3)}$	$809.1_{(0.3)}$	$864.6_{(0.3)}$	$829.6_{(0.3)}$	$854.0_{(0.3)}$
7,7	$852.0_{(0.4)}$	$847.1_{(0.4)}$	$855.7_{(0.4)}$	$855.0_{(0.4)}$	$853.3_{(0.4)}$
7,10	$877.9_{(0.3)}$	$862.6_{(0.4)}$	$896.7_{(0.3)}$	$875.5_{(0.3)}$	$883.5_{(0.4)}$

Table 9 shows the estimated Kullback-Leibler divergence of the nominal p-value distribution under the null hypothesis of each test compared to the theoretical uniform distribution, from a two-sample comparison design, with varying sample sizes. One can see that both methods based on our decorrelated and reduced dataset, *i.e.*, the reduced Levene's test ("RLev") and MRPP, have small Kullback-Leibler distances under all sample sizes. The original Levene's test ("Lev") generally has a poor null distribution. The F-test approximates the null well when sample sizes are small, but not when sample sizes increase. The Brown-Forsythe test ("BF") has a close to uniform null distribution when both sample sizes are even, but it becomes worse if one of the sample sizes is odd, and is usually the worst compared to other methods when both sample sizes are odd. This is because the definition of median depends on the parity of sample size. In terms of null distribution, our proposed RLev and MRPP outperform other methods under our realistic simulation settings.

Table 10 shows the probability that the test statistic under the alternative is at least as extreme as the test statistic under the null when $\mathrm{E}(r_g) = 1$. This corresponds to small departure of the alternative from the null. We can see that whenever the two sample sizes are the same, all five tests have similar power. Otherwise, the F-test, the modified MRPP, and the reduced Levene's test have better power than the original Levene's test and the Brown-Forsythe test.

Table 11 shows the same probability when $\mathrm{E}(r_g) = 10$, which corresponds to a large difference between the null and the alternative. In this case, the reduced Levene's test does not perform well, but the Brown-Forsythe test is often the best, usually closely followed by Levene's test and the modified MRPP. The performance of the F-test seems to depend on the balancedness of the design—when sample sizes are

equal, its power is high; but when sample sizes are different, its power deteriorates greatly.

Considering both small and large departures from the null hypotheses, the modified MRPP is the only one that always has relatively good power. Since it also controls type I error better than others (Table 9) and directly applies to any high dimensional situation, it is our recommended test for heterogeneity in practice to replace the aforementioned alternatives.

4. SUMMARY

In this study, we proposed several new model selection procedures for selecting the variance part of linear models. Our modified MRPP test has tighter control of type I errors and also has good power compared to other methods. Our cross-validation procedures are able to differentiate linear models that only differ in the variance assumptions. We also give the correct AICc formula that removes the bias in AIC when multiple variances need to be estimated independently.

Through real data based simulation, we found that using the correct models does not necessarily provide the best separation between differentially and equivalently expressed genes, although using the correct models can control FDR at desired levels. A hybrid procedure to decouple FDR control and differential expression detection is recommended, as in [12].

Variance model selection for mixed linear models is more complicated, primarily because the estimated variance components may lie on the boundary of the parameter space with positive probability [16]. We may envisage using some of our methods, *e.g.*, AICc, in special types of mixed models, but it would be valuable in future studies to extend our methods to general mixed models.

Our simulations suggest that neither shrinkage estimation within a dataset nor model selection across several datasets is always preferred to the other. Fortunately, it is straightforward to combine both approaches by first pooling across datasets (if our proposed methods suggest that pooling will be beneficial) and then shrinking the resulting estimates by borrowing information across genes using a procedure like *limma*.

In summary, for microarray data analysis, our general recommendation for ranking genes is to use a pooled analysis only when variances are judged to be equal or when variances differ but the contrasts of interest are within the datasets with larger variance. For control of FDR, pooled analysis should only be applied when variances are judged to be equal and the combined sample size is moderately large, irrespective of which analysis has been used to rank genes. For both ranking genes and controlling FDR, shrinkage estimation of variances across genes is recommended, irrespective of whether additional datasets will be used to estimate variances. If homogeneity of variances will be tested separately for each gene, then the modified MRPP procedure is preferred to other univariate tests.

Received 21 January 2010

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceeding of the Second International Symposium of Information Theory*, B. Petrov and F. Caski, Eds. Springer-Verlag, New York, Akademiai Kiado, Budapest, 267–281. MR0483125

[2] Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19**, 6, 716–723. MR0423716

[3] Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics* **16**, 1, 125–127. MR0343481

[4] Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 6, 509–519.

[5] Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles—database and tools. *Nucl. Acids Res.* **33**, suppl_1, D562–566.

[6] Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **160**, 901, 268–282.

[7] 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. Series B (Methodological)* **57**, 1, 289–300. MR1325392

[8] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 3, 345–370. MR0914460

[9] Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association* **69**, 346, 364–367.

[10] Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York. MR1919620

[11] Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 1, 59–75.

[12] Demirkale, C. Y., Nettleton, D., and Maiti, T. (2010). Linear mixed model selection for false discovery rate control in microarray data analysis. *Biometrics* **66**, 2, 621–629.

[13] Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* **18**, suppl_1, S105–110.

[14] Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 456, 1151–1160. MR1946571

[15] Geller, S. C., Gregg, J. P., Hagerman, P., and Rocke, D. M. (2003). Transformation and normalization of oligonucleotide microarray data. *Bioinformatics* **19**, 14, 1817–1823.

[16] Greven, S. and Kneib, T. (2009). On the behavior of marginal and conditional Akaike information criteria in linear mixed models. Tech. Rep. Working Paper 179, Johns Hopkins University, Dept. of Biostatistics.

[17] Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)* **41**, 2, 190–195. MR0547244

[18] Hannenhalli, S., Putt, M. E., Gilmore, J. M., Wang, J., Parmacek, M. S., Epstein, J. A., Morrisey, E. E., Margulies, K. B., and Cappola, T. P. (2006). Transcriptional genomics associates FOX transcription factors with human heart failure. *Circulation* **114**, 12, 1269–1276.

[19] Hartley, H. O. (1950). The use of range in analysis of variance. *Biometrika* **37**, 3/4, 271–280. MR0039958

[20] Huang, X. and Pan, W. (2002). Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. *Functional and Integrative Genomics* **2**, 3, 126–133.

[21] Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 2, 297–307. MR1016020

[22] Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M., and Lee, J. K. (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **19**, 15, 1945–1951.

[23] Kamb, A. and Ramaswami, M. (2001). A simple method for statistical analysis of intensity differences in microarray-derived gene expression data. *BMC Biotechnol 1*, 8.

[24] Levene, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.*, I. Olkin, Ed. Stanford University Press, Palo Alto, CA, 278–292. MR0120709

[25] Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica 12*, 31–46. MR1894187

[26] Mallows, C. L. (1973). Some comments on C_P. *Technometrics* **15**, 4, 661–675.

[27] Mielke, P. W. and Berry, K. J. (2007). *Permutation methods: a distance function approach*, 2nd ed. Springer, New York. MR2378190

[28] Nettleton, D., Recknor, J., and Reecy, J. M. (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics* **24**, 2, 192–201.

[29] Patterson, H. D. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 3, 545–554. MR0319325

[30] R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[31] SAS Institute, Inc. (2008). *SAS/STAT(R) 9.2 User's Guide*. SAS Institute, Inc.

[32] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 2, 461–464. MR0468014

[33] Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics* **8**, 1, 147–164. MR0557560

[34] Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 1, 45–54. MR0614940

[35] Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 20, 3940–3941.

[36] Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, 1, Article 3. MR2101454

[37] Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods*, 8th ed. Iowa State University Press, Ames, IA. MR1017246

[38] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 4, 583–639. MR1979380

[39] Spira, A., Beane, J. E., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y. M., Calner, P., Sebastiani, P., Sridhar, S., Beamis, J., Lamb, C., Anderson, T., Gerry, N., Keane, J., Lenburg, M. E., and Brody, J. S. (2007). Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med* **13**, 3, 361–366.

[40] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 16, 9440–9445. MR1994856

[41] Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences) 153*, 12–18.

[42] Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., and Shaughnessy, J. D., J. (2003). The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. *N Engl J Med* **349**, 26, 2483–2494.

[43] Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. *Journal of the American Statistical Association 102*, 113–122. MR2293304

[44] Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 9, 5116–5121.

[45] Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 6, 625–637.

Long Qu
Iowa State University
Department of Statistics and Department of Animal Science
229 Kildee Hall
Ames, IA 50011-3150
E-mail address: longor@iastate.edu

Dan Nettleton
Iowa State University
Department of Statistics
2115 Snedecor Hall
Ames, IA 50011-1210
E-mail address: dnett@iastate.edu

Jack C. M. Dekkers
Iowa State University
Department of Animal Science
239D Kildee Hall
Ames, IA 50011-3150
E-mail address: jdekkers@iastate.edu

Nicola Bacciu
INRA, GARen, Agrocampus
35000 Rennes, France
E-mail address: Nicola.Bacciu@rennes.inra.fr