

# Gene Expression Analysis with the Parametric Bootstrap

BY MARK J. VAN DER LAAN AND JENNY BRYAN

*Division of Biostatistics, University of California  
Earl Warren Hall 7360, Berkeley, California 94720-7360  
e-mail: laan@stat.berkeley.edu*

## SUMMARY

Recent developments in microarray technology make it possible to capture the gene expression profiles for thousands of genes at once. With this data researchers are tackling problems ranging from the identification of “cancer genes” to the formidable task of adding functional annotations to our rapidly-growing gene databases. Specific research questions suggest patterns of gene expression that are interesting and informative, for instance, genes with large variance or groups of genes that are highly correlated. Cluster analysis and related techniques are proving to be very useful. However, such exploratory methods alone do not provide the opportunity to engage in *statistical inference*. Given the high-dimensionality (thousands of genes) and small sample sizes (often  $< 30$ ) encountered in these datasets, an honest assessment of sampling variability is crucial and can prevent the over-interpretation of spurious results. We describe a statistical framework that encompasses many of the analytical goals in gene expression analysis; our framework is completely compatible with many of the current approaches and, in fact, can increase their utility. We propose the use of a deterministic rule, applied to the parameters of the gene expression distribution, to select a target subset of genes that are of biological interest. In addition to subset membership, the target subset can include information about relationships between genes, such as clustering. This target subset presents an interesting parameter that we can estimate by applying the rule to the sample statistics of microarray data. The parametric bootstrap, based on a multivariate normal model, is used to estimate the distribution of these estimated subsets and relevant summary measures of this sampling distribution are proposed. We focus on rules that operate on the mean and covariance. Using Bernstein’s Inequality, we obtain consistency of the subset estimates, under the assumption that the sample size converges faster to infinity than the logarithm of the number of genes. We also provide a conservative sample size formula guaranteeing that the sample mean and sample covariance matrix are *uniformly* within a distance  $\epsilon > 0$  of the population mean and covariance. The practical performance of the method using a cluster-based subset rule is illustrated with a simulation study. The method is illustrated with an analysis of a publicly available leukemia data set.

*Some key words:* Gene expression, multivariate normal, parametric bootstrap, cluster analysis.

<sup>0</sup>This research has been supported by NIAID grant 1R01 AI46182-01.

## 1. INTRODUCTION

1.1. *Microarray context*

Microarray studies are swiftly becoming a very significant and prevalent tool in biomedical research. The microarray technology allows researchers to monitor the expression of thousands of genes simultaneously. A readable introduction to microarrays can be found in Marshall (1999) and a more technical overview is given in the “The Chipping Forecast” (1999).

By comparing gene expression profiles across cells that are at different stages in some process, in distinct pathological states, or under different experimental conditions, researchers gain insight into the roles and reactions of various genes. For example, one can compare healthy cells to cancerous cells within subjects in order to learn which genes tend to be over (or under) expressed in the diseased cells; regulation of such genes could produce effective cancer treatment and/or prophylaxis. DeRisi et al. (1996) suppressed the tumorigenic properties of human melanoma cells and compared gene expression profiles among “normal” and modified melanoma cells; this experiment allowed investigators to study the differential gene expression that is associated with tumor suppression. Data analysis methods appropriate for microarray data are surveyed by Claverie (1999), Eisen et al. (1998), and Herwig et al. (1999).

Recent microarray studies have relied heavily on clustering procedures. Eisen et al. (1998) apply a hierarchical cluster analysis algorithm to an empirical correlation matrix and Golub et al. (1999) use a neural network algorithm called self-organizing maps (SOM) which, like K-means clustering and the partitioning around medoids (PAM) of Kaufman and Rousseeuw (1990), places objects into a fixed number of clusters. We feel that such approaches suffer from two deficiencies, which we address in this paper. First, since these techniques are used in a purely data exploratory manner, they lack important notions such as parameter, parameter estimate, consistency and confidence. Second, techniques that are purely descriptive and *ad hoc* make it difficult to design a study to meet particular goals.

1.2. *Overview of the statistical method*

We begin by describing a basic microarray experiment, resulting in a relative gene expression profile for one experimental unit. Two tissue or cell line samples are collected and the mRNA of each is labelled with a different red or green dye. Subsequently, the mRNA samples are combined and washed over a microarray prepared with the cDNA of  $p$  genes. The labelled RNA of gene  $j$  will attach to the corresponding spot on the microarray,  $j = 1, \dots, p$ . A scanner measures the red and green intensity at each of the spots. A natural summary measure is the ratio of these gene-specific red and green intensities. Typically, one of the samples is a control and the ratios are defined as test over control. When component  $X_j > 1$ , the DNA of gene  $j$  has a higher expression in the test sample than in the control sample. Let  $\mathbf{X}$  be the  $p$ -dimensional column vector of ratios representing the relative gene expression profile for a subject or cell line randomly drawn from a well-defined population. Suppose that we observe  $n$  i.i.d. copies  $\mathbf{X}_1, \dots, \mathbf{X}_n$  of this random vector  $\mathbf{X}$ , for example, once for each of  $n$  individuals.

For clarity, consider two concrete examples: In the dataset that originally motivated this work, the population of interest is human colon cancer patients and for each subject we have a sample of healthy colon tissue (control) and colon tumor tissue (test). From such data, we want to find a subset of genes for which differential expression is associated with cancer. In the dataset we analyze in section 5, the population of interest is human acute leukemia patients described by Golub et al. (1999). The data actually arise from a microarray technology slightly different than

the cDNA arrays described above, namely, an oligonucleotide array produced by Affymetrix. With this technology, measurements of absolute expression levels are obtained; therefore, for each gene we obtain a positive number reflecting expression level instead of the ratio produced in a cDNA experiment. In any case, the dataset contains expression profiles for patients with two distinct types of leukemia. Our data analysis in section 5 focuses on finding genes whose expression best distinguishes the two tumor classes and are, therefore, useful in diagnosis.

In light of the typical scientific goals, we generally wish to find (1) genes that are *differentially expressed*, that is, expression is different in the test sample relative to the control, and (2) groups of genes which are *significantly correlated with each other*. We are interested in genes whose expression levels tend to vary together, because such genes might be part of the same causal mechanism.

Since k-fold over-expression represents the opposite of k-fold under-expression, it is natural to use a logarithmic transformation: let  $Y_j^* = \log(X_j)$   $j = 1, \dots, p$ . In addition, to control the effect of outliers and to obtain nonparametric consistency results proved later in this paper, we also propose to truncate the log-ratios by a user-supplied constant  $M$ :

$$Y_j = \begin{cases} Y_j^* & \text{if } |Y_j^*| < M, \\ M \times \text{sgn}(Y_j^*) & \text{if } |Y_j^*| \geq M. \end{cases} \quad j = 1, \dots, p, \quad 0 < M < \infty.$$

Another truncation approach would be to examine centered data  $Y_j^* - \hat{\mu}_j^*$  and truncate all observations that were, for example, greater than 3 standard deviations in absolute value. One could also assume a multivariate normal model for the raw log-ratios and rely on the light tails of the normal distribution for consistency results. Let  $\mathbf{Y}$  be the column vector with component  $j$  being equal to  $Y_j$ ,  $j = 1, \dots, p$ . Denote the expectation, covariance, and correlation of  $\mathbf{Y}$  by  $\boldsymbol{\mu}$ ,  $\boldsymbol{\Sigma}$ , and  $\boldsymbol{\rho}$ , respectively.

Suppose that we know  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ , and that subject matter experts believe that certain patterns of gene expression distinguish specific genes as important. Then a natural question is “How should we select a subset  $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow \mathbf{S}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  of genes that merit special attention?” We might also wish to regard  $\mathbf{S}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  as a set of genes that is subdivided into several groups labeled from 1 to  $K$ . We can identify such a subset  $\mathbf{S}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  by a  $p$ -vector  $\mathbf{S}$  whose components take values in  $\{0, \dots, K\}$ . If  $S_j = 0$ , then gene  $j$  is excluded from the subset and if  $S_j = k$ ,  $k \in \{1, \dots, K\}$ , then gene  $j$  is included in the subset and carries label  $k$ . At times, we will also describe the subset as a set of gene indices  $j$ ,  $j \in \{1, \dots, p\}$ ; this is equivalent to setting  $S_j = 0$  for genes not in the subset and to some integer between 1 and  $K$  otherwise. Hereinafter  $\mathcal{S} \equiv \mathbf{S}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  will represent the target subset of genes that we wish to distinguish as important.

As an example of a very simple rule, one could define  $\mathbf{S}(\cdot, \cdot)$  as  $\{j : \mu_j > C\}$ , for some  $C > 0$ . A more sophisticated subset rule would be to (1) select those genes which are at least 3-fold differentially expressed w.r.t. the geometric mean (i.e. only include gene  $j$  if  $|\mu_j| > \log 3$ ); and (2) construct a correlation-distance matrix for these differentially expressed genes from the appropriate elements of  $\boldsymbol{\rho}$ ; and (3) apply a clustering algorithm to (some function of) this distance-matrix; and possibly (4) only include those genes in  $\mathbf{S}$  that are closest to the cluster centers. In fact, most of the analytical techniques currently being applied to gene expression data (for example Eisen et al., 1998; Golub et al., 1999) operate on the mean and covariance and are, therefore, perfect candidates for the type of subset rule proposed in this paper. It is not necessary for the subset rule to eliminate genes at all, although it generally advantageous to do so. Even if the rule simply applies labels that have a stable meaning – for example, by employing a supervised clustering technique that finds clusters around pre-specified genes – the methods we propose would allow the analyst to assess the stability of these clusters.

Given a well-defined subset rule  $\mathbf{S}(\cdot, \cdot)$ , a natural estimate of the target subset  $\mathcal{S}$  is  $\hat{\mathbf{S}}_n \equiv \mathbf{S}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ , where  $\hat{\boldsymbol{\mu}}_n$  and  $\hat{\boldsymbol{\Sigma}}_n$  are the sample mean and covariance, respectively, of the (truncated) data. We prove the consistency of  $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$  and  $\mathbf{S}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$  (see section 3) *nonparametrically* when  $n/\log(p(n)) \rightarrow \infty$  and  $M < \infty$ . The case where  $p = \infty$  and  $p \gg n$  is extremely relevant, as microarray experiments already produce data on 20000 genes and, in the future, we will encounter datasets with all human genes (estimated to be between 35000 and 140000). In stark contrast, sample sizes often fall below 30. We also provide a nonparametric sample size formula that guarantees with probability at least  $0 < \delta < 1$  that the maximal difference between  $\hat{\boldsymbol{\mu}}_n$  and  $\boldsymbol{\mu}$  is smaller than  $\epsilon$  and similarly for  $\hat{\boldsymbol{\Sigma}}_n$  and  $\boldsymbol{\Sigma}$ . If one is willing to assume that  $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  has a multivariate normal distribution, then the truncation is not needed, but in this paper we aim to be as nonparametric as possible.

The sampling distribution of the estimated subsets  $\hat{\mathbf{S}}_n$  provides valuable information for the analyst. One might wish to choose the sample size and/or subset rule in order to ensure the reproducibility of certain results or to realize some other performance measure. As an example of a feature we would hope to see reproduced in samples, consider a gene  $j$  that appears in  $\mathcal{S}$ . For a particular data-generating distribution, sample size  $n$ , and subset rule  $\mathbf{S}(\cdot, \cdot)$ , there is a probability  $p_j$  that gene  $j$  will appear in the estimated subset  $\hat{\mathbf{S}}_n$  produced by a randomly drawn sample; we will call such probabilities  $p_j$  “single-gene probabilities”. If the single-gene probabilities are low for many of the genes in  $\mathcal{S}$ , we might choose to increase the sample size or select a subset rule that is easier to estimate. If the single-gene probabilities are generally high, we might proceed with the study and, when we observe estimates of  $p_j$  that are close to 1, feel confident that those genes are in  $\mathcal{S}$ .

Since we want to determine the membership of a specific set, it is natural to apply conventional measures of test quality, such as sensitivity and positive predictive value, to any procedure we devise. In this context, sensitivity is the proportion of the target subset that also falls in the estimated subset and positive predictive value is the proportion of the estimated subset that is also in the target subset.

Determining single-gene probabilities and the distribution of subset quality measures requires knowledge of the actual sampling distribution of  $\hat{\mathbf{S}}_n$ . In order to estimate these quantities we use the parametric bootstrap. One reason we prefer a parametric over a nonparametric bootstrap is the fact that the parametric bootstrap is asymptotically valid under relatively mild conditions compared to those required by the nonparametric bootstrap (Giné and Zinn, 1990). In general, the asymptotic validity of the parametric bootstrap requires that the chosen parametric model be correct. However, as long as we choose a parametric model that places no constraints on  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , even when it is incorrect, the parametric bootstrap will still consistently estimate the degenerate limit distribution of  $\hat{\mathbf{S}}$  and  $\sqrt{n}(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu})$ . See 1.3 for more discussion. Specifically, we use the multivariate normal model  $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  and, based on the data we have seen, we believe this to be a reasonable choice, after truncation.

The bootstrap (Efron and Tibshirani, 1993) was first used to investigate the reproducibility of certain features of phylogenetic trees by Felsenstein (1985). Efron and Tibshirani (1998) later took up this problem more generally and termed it the “problem of regions”. They ask: given an interesting feature in an observed descriptive statistic, how confident can we be that this feature is present in the data-generating distribution? Efron and Tibshirani also link this confidence measure, in certain settings, to frequentist  $p$ -values and Bayesian a posteriori probabilities.

### 1.3. Relevance of the multivariate normality assumption

Because  $\widehat{\mathbf{S}}_n$  is only a function of the sample mean and covariance, resampling from a distribution with consistent first and second central moments, such as the parametric bootstrap described below, consistently estimates the degenerate limit-distribution at  $\mathcal{S}$ , even when the multivariate normality assumption is violated. This is another good reason for choosing the multivariate normal as our model to carry out the parametric bootstrap.

By the central limit theorem, for finite  $p$  (we also prove a CLT for  $p = \infty$ ),  $n$  large, and any data generating distribution of  $\mathbf{Y}$ , we know that the distribution of  $\sqrt{n}((\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}), (\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}))$  is multivariate normal. The normal limiting distribution  $N(\mathbf{0}, \boldsymbol{\Sigma})$  of  $\sqrt{n}(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu})$  only depends on the covariance matrix  $\boldsymbol{\Sigma}$  of  $\mathbf{Y}$ , which is consistently estimated. As a consequence, though our multivariate truncated normality assumption helps us to carry out a parametric bootstrap procedure, the consistency of this parametric bootstrap for estimating the distribution of  $\sqrt{n}(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu})$  as  $n \rightarrow \infty$  does not rely on it. This is important since  $\boldsymbol{\mu}$  is used in the first stage of most subset rules and, in fact, a subset rule *only* based on  $\boldsymbol{\mu}$  is already practically meaningful. However, because the covariance matrix of the normal limiting distribution of  $\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})$  is not identified by  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , the consistency of the parametric bootstrap for estimating the distribution of  $\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})$  as  $n \rightarrow \infty$  does rely on the multivariate normality assumption. To summarize, our estimators of  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  are nonparametrically consistent and the parametric bootstrap relies on the multivariate normality assumption only for estimating the distribution of  $\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})$ .

If we truncate the log-ratios as proposed above, then the multivariate normal model cannot be true. In van der Laan and Bryan (2000) this minor inconsistency is handled by assuming the multivariate normal model on the untruncated log-ratios  $\mathbf{Y}^*$  and making the truncation part of the data generating mechanism (and thus the parametric bootstrap). However, by the arguments given above, this modification is hardly worthwhile, while it makes the notational setting much more complex.

### 1.4. Organization of the paper

In section 2 we describe subset rules and the parametric bootstrap method for estimating the single-gene probabilities, sensitivity and positive predictive value of  $\widehat{\mathbf{S}}_n$ .

Because  $p$  is typically very large relative to  $n$ , in section 3 we describe the performance of the estimators and parametric bootstrap when  $p = p(n) \gg n$ . We prove that if  $n/\log(p(n))$  converges to infinity, then  $(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n) - (\boldsymbol{\mu}, \boldsymbol{\Sigma})$  converges uniformly to zero (Theorem 3.1). For certain subset rules, we prove that  $\widehat{\mathbf{S}}_n$  equals  $\mathcal{S}$  with probability tending to one and the bootstrap estimate of the distribution of  $\widehat{\mathbf{S}}_n$  converges to the degenerate distribution at  $\mathcal{S}$  (Theorems 3.3 and 3.4). We also provide a sample size formula (Theorem 3.2).

In section 4 we describe a simulation study illustrating the performance of our proposed methodology. In section 5 we analyze a publicly available data set in human acute leukemia.

## 2. THE ESTIMATED SUBSET AND THE PARAMETRIC BOOTSTRAP

### 2.1. Subset rules

We propose several simple, but easily interpretable, subset rules and all are simply functions of the parameters  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . We have found it natural to divide the subset rule into three phases: (1) a pre-screen in which certain genes are eliminated; (2) a mid-rule in which inter-relationships

Table 1. *Subset rule examples.*

Pre-screen	Distance metric	Mid-rule	Post-screen
$\mu_j > \log(\delta_1)$	Euclidean distance	PAM	$D_{ij} < \delta_2$ ,
$ \mu_j  > \log(\delta_1)$	$1 -  \rho_{ij} , 1 - \rho_{ij}$	PAM, with fixed medoids	for some cluster center $i$
$\mu_j a >$	1 - Eisen's modified	K-means clustering	$silhouette_j > \delta_2$
$A + \sigma_j \Phi^{-1}(q)$	correlation	Hierarchical clustering	(part of PAM output)

between genes are sought; and (3) a post-screen in which even more genes are eliminated. We emphasize that it is not necessary to employ all three phases of the rule and, therefore, a clustering algorithm alone can be regarded as an example of such a rule, whenever the distance metric is a function of  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . For example, this is the case with Euclidean distance, correlation distance, the modified correlation distance proposed by Eisen et al. (1998), and principal component based metrics. From now on, we denote the distance between genes  $i$  and  $j$  by  $D_{ij}$ , the  $p$  by  $p$  symmetric matrix of such distances by  $\mathbf{D}$ , and we assume that  $\mathbf{D}$  is determined by  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .

Table 1 presents examples of the rules and metrics one can work with. A common requirement for inclusion in the subset is *differential expression* and we use the pre-screen to retain only those with sufficient evidence of differential expression. The table presents pre-screens that range from very simple cutoffs to those that determine whether a certain proportion  $q$  of the population exhibits a sufficient level  $A$  of over and/or under (determines  $a$ ) expression. We then seek groups of genes that tend to be coexpressed; a clustering algorithm, such as PAM (Kaufman and Rousseeuw, 1990, chap. 2), is a typical mid-rule, but many other clustering and neural network algorithms are also suitable. Finally, since clustering algorithms place all objects into clusters, even if there is little evidence to favor one cluster assignment over another, we often use the post-screen to retain only those genes that appear to be well-matched to their cluster. One could use actual distances to cluster centers or members to make this determination or, as in the case of the “silhouettes” in PAM, there may be other useful output from the clustering procedure one can exploit.

To be concrete, let us define a particular rule which will be revisited later. We retain genes that have a mean greater than a constant  $C_\mu > 0$  and that have a correlation coefficient with another included gene that is greater than a constant  $0 < C_\rho < 1$ :

$$\mathbf{S}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{j : \mu_j, \mu_i > C_\mu, \rho_{ij} > C_\rho, \text{ for some } j \neq i\}, \quad (1)$$

This rule seeks genes that are over expressed and that have a large correlation with at least one other over expressed gene.

## 2.2. The parametric bootstrap.

Our goal is to estimate the distribution of  $\hat{\mathbf{S}}_n \equiv \mathbf{S}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) \in \{0, \dots, K\}^p$ , where  $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$  are the observed mean and covariance matrix of a size  $n$  sample from a  $N_{p(M)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  distribution, where we will treat  $K$  as fixed. The parametric bootstrap described below could also be used to address uncertainty of a data adaptively determined  $K$ , but in this paper we want to focus on the “fixed  $K$ ” case. The parametric bootstrap estimates the distribution of  $\hat{\mathbf{S}}_n$  with the distribution of  $\tilde{\mathbf{S}}_n \equiv \mathbf{S}(\tilde{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\Sigma}}_n)$ , where  $(\tilde{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\Sigma}}_n)$  are the observed mean and covariance matrix of a size  $n$  sample from a  $N_p(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ . From this point on, sample quantities (first-generation draws from the data-generating distribution) will be indicated by hats and bootstrap quantities (second-generation draws from statistics of an observed first-generation sample) with tildes.

When we draw from a  $N_p(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ , we will be faced with a singular covariance matrix  $\hat{\boldsymbol{\Sigma}}_n$  when  $n$  is smaller than  $p$ . In that case we add to the diagonal elements of  $\hat{\boldsymbol{\Sigma}}_n$  an arbitrarily small number  $\lambda > 0$ , which produces a nonsingular covariance matrix that is extremely close to  $\hat{\boldsymbol{\Sigma}}_n$ . This ensures that we are sampling from a well-defined distribution.

So, for  $b = 1, \dots, B$  ( $B$  large), we draw  $n$  observations (i.e. microarrays)  $\tilde{\mathbf{Y}}_1^b, \dots, \tilde{\mathbf{Y}}_n^b$  from  $N_p(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ , compute the observed statistics  $(\tilde{\boldsymbol{\mu}}_n^b, \tilde{\boldsymbol{\Sigma}}_n^b)$ , and record the estimated bootstrap subset  $\tilde{\mathbf{S}}_n^b = \mathbf{S}(\tilde{\boldsymbol{\mu}}_n^b, \tilde{\boldsymbol{\Sigma}}_n^b)$ . This provides us with  $B$  realizations  $\tilde{\mathbf{S}}_n^1, \dots, \tilde{\mathbf{S}}_n^B$  of  $\tilde{\mathbf{S}}_n = \mathbf{S}(\tilde{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\Sigma}}_n)$ . We use this observed distribution as an estimate of the distribution of  $\tilde{\mathbf{S}}_n$  and, for  $n$  large enough, one can view the sampling distribution of  $\tilde{\mathbf{S}}_n$  as an estimate of the distribution of  $\hat{\mathbf{S}}_n = \mathbf{S}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ .

### 2.2.1. Important features of the sampling distribution.

To be specific, consider one of the cluster-based subset rules. Because of the high dimensionality of  $\hat{\mathbf{S}}_n$ , we limit our focus to certain aspects of the empirical distribution of  $\hat{\mathbf{S}}_n$ . Note that the values and/or distributions of the quantities defined in this section certainly depend on the sample size  $n$ ; we will employ notation with and without the  $n$ , depending on the context. Indicate the size of a set  $\mathcal{A}$  by  $|\mathcal{A}|$ . Let

$$\begin{aligned} p_j &= p_{j,n} = P(\hat{S}_j > 0) \\ P_{ij} &= P_{ij,n} = P(\hat{S}_i > 0, \hat{S}_j > 0) \\ Q_{ij} &= Q_{ij,n} = P(\hat{S}_i = \hat{S}_j > 0) \\ sens &= sens_n = |\mathcal{S} \cap \hat{\mathbf{S}}|/|\mathcal{S}| \\ ppv &= ppv_n = |\mathcal{S} \cap \hat{\mathbf{S}}|/|\hat{\mathbf{S}}| \end{aligned}$$

The quantities  $p_j$ ,  $P_{ij}$ , and  $Q_{ij}$  are referred to collectively as “feature-specific probabilities”. The random variables  $sens$  and  $ppv$  are called “quality measures”. It is important to note that the concepts of sensitivity and predictive value as employed here differ from their epidemiological counterparts, in that the  $p$  genes are not assumed to be i.i.d. Since these quantities are functions of the estimated subset, the distribution of sensitivity and positive predictive value are to be considered when evaluating a proposed subset rule.

The significance of sensitivity is rather obvious, but we would like to emphasize the importance of positive predictive value as well. If scientists use the estimated subset  $\hat{\mathbf{S}}_n$  as a means for selecting a relatively small set of genes for intensive study, it is crucial that the predictive value be high, since a great deal of time and money could be wasted otherwise. This is especially relevant when  $p$  is very large. Since an estimated subset for which 50% of the genes are false positives might not be considered usable, information on the predictive value of the estimated subset could alert researchers to the need to collect more data or to choose a different subset rule.

The bootstrap analogue of the above feature-specific probabilities is the empirical frequency in the bootstrap replicates of the appropriate event. Likewise the bootstrap estimate of the distribution of a quality measure will be based on the appropriate empirical proportions. For example,  $\hat{p}_j = \hat{p}_{j,n} = \frac{1}{B} \sum_b I(\hat{S}_j^b > 0)$  and  $\widehat{sens} = \widehat{sens}_n^b = |\hat{\mathbf{S}} \cap \tilde{\mathbf{S}}^b|/|\hat{\mathbf{S}}|$ . All of these quantities are retained in the bootstrap. For practical reasons, we focus on the single-gene probabilities,  $p_j$ , and sort the genes in descending order based on this. We report the top-ranked genes and ensure that all members of  $\hat{\mathbf{S}}_n$  are included. In such a list, the genes which fall “deep” into the target subset  $\mathcal{S}$  will typically appear before the genes which barely qualified for inclusion into  $\mathcal{S}$ . The scientist can begin carefully investigating the top ranked genes.

In some settings, there may be a subset of genes that are not only excluded from the target subset  $\mathcal{S}$ , but are regarded as particularly unsuitable for further study. The definition of such genes is completely up to the user, but will generally correspond to genes that lie *far* outside the target subset. We will denote this subset by  $\mathcal{L}$ , which is a subset of  $\mathcal{S}^c$ , the complement of  $\mathcal{S}$ . For example, one might regard the set  $\mathcal{L} = \{j : |\mu_j| < D_\mu\}$ , where  $D_\mu < C_\mu$ , as particularly inappropriate for further study. The proportion of such genes in the estimated subset is of great interest; we will refer to this quantity as the “proportion of extremely false positives” or *pefp*. If this proportion is always very low, one can be reasonably confident that the top ranked genes of the reported subset contain no extremely false positives. We define  $\xi_j = \xi_j(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 1$  if gene  $j$  is in  $\mathcal{L}$  and  $\xi_j = 0$  otherwise,  $j = 1, \dots, p$ . Now, the proportion of extremely false positives (*pefp*) for the estimated subset  $\widehat{\mathbf{S}}_n$  can be defined as:

$$pefp = pefp_n = \frac{1}{|\widehat{\mathbf{S}}_n|} \sum_{j=1}^p I(\xi_j = 1, \widehat{S}_j > 0).$$

As with sensitivity and predictive value, we can use the parametric bootstrap to estimate the expectation  $E(pefp)$  and variance  $\text{Var}(pefp)$ . Note that if  $\mathcal{L} = \mathcal{S}^c$ , *pefp* is simply one minus the positive predictive value. Therefore, *pefp* becomes interesting only when  $\mathcal{L}$  is considerably smaller than  $\mathcal{S}^c$ .

Other important quantities of interest are the 0.95-quantiles of  $|\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}| \equiv \max_j |\widehat{\mu}_{nj} - \mu_j|$  and  $\max_{i,j} |\widehat{\boldsymbol{\rho}}_n - \boldsymbol{\rho}|$ . It should also be noted that one could replace the PAM procedure described above with some “supervised” clustering method that allows the analyst to specify the cluster centers. If the centers were fixed at genes of known function, the clusters have a coherent meaning throughout the bootstrap. In that case, we can also track how often each gene appears in each fixed-center cluster and, thereby, obtain information on cluster stability.

### 2.2.2. Interpretation of the output of the parametric bootstrap.

By relying on asymptotic properties established in Section 3, we can view the relative frequencies  $(\widehat{p}_j, \widehat{P}_{ij}, \widehat{Q}_{ij})$  as estimates of the probabilities  $(p_j, P_{ij}, Q_{ij})$ . Consider now the situation in which  $n$  is too small to reasonably assume that  $(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)$  is close to  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . In this case, it does not follow that the distribution of  $\widetilde{\mathbf{S}}_n$  is close to the distribution of  $\widehat{\mathbf{S}}_n$ . It is our experience that the results of the parametric bootstrap are still valuable. One can simply interpret the results as a simulation study for estimation of  $\widehat{\mathbf{S}}_n$  when sampling from  $N(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)$ . Findings of such a simulation study (such as a low predictive value) will demonstrate the difficulty of estimating  $\mathcal{S}$  with the given sample size  $n$ . In particular, one can run the parametric bootstrap for several subset rules and thereby determine which types of subsets can be reasonably estimated with the available sample size.

### 2.2.3. Finite sample bias of single-gene probabilities.

Suppose for a moment that we are not concerned with distinctions within the target subset and we can reduce  $\mathcal{S}$  to a  $p$ -dimensional vector with  $\{0, 1\}$ -valued components  $\delta_j = I(j \in \mathcal{S})$ ,  $j = 1, \dots, p$ . Make the same simplification for  $\widehat{\mathbf{S}}_n$ . We have that  $\widehat{\mathbf{S}}_n$  is an estimator of  $\mathcal{S}$ . Note that  $E(\widehat{S}_j) = p_j$ , for  $j = 1, \dots, p$ , which shows that, if  $\delta_j = 1$ , then  $\widehat{\mathbf{S}}_n$  is biased low and if  $\delta_j = 0$ , then  $\widehat{\mathbf{S}}_n$  is biased high. Therefore,  $\widehat{\mathbf{S}}_n$  is an asymptotically consistent estimator of  $\mathcal{S}$  with a well understood bias for finite samples. The same argument holds for the conditional bias of the bootstrap subset  $\widetilde{\mathbf{S}}_n$  for the estimated subset  $\widehat{\mathbf{S}}_n$  given  $(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)$ . This has important

implications for the parametric bootstrap. We use  $\hat{p}_j$  to estimate  $p_j$  and, as a consequence of the facts explained above, it is biased in finite samples. Informally, if  $j$  is “deep” into  $\mathcal{S}$ , then  $p_j$  will be greater than 0.5 and  $\hat{p}_j$  will tend to be biased low. Conversely, if  $j$  is “far” from  $\mathcal{S}$ , then  $p_j$  will be less than 0.5 and  $\hat{p}_j$  will tend to be biased high. The importance of “deep” and “far” in the above argument decreases as  $n$  grows large. In van der Laan and Bryan (2000) we provide an algebraic argument for this bias direction.

### 3. ASYMPTOTIC THEORY.

Our proposed method for analyzing gene expression data reports an estimated subset  $\hat{\mathbf{S}}_n \equiv \mathbf{S}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$  and bootstrap estimates of the feature-specific probabilities and other quality measures. In this section we prove the consistency of  $\hat{\mathbf{S}}_n$  and the bootstrap estimate of its distribution under appropriate conditions. Additionally, we provide a sample size formula that controls the probability of the estimated subset containing any extremely false positives.

Because  $p$  is typically much larger than  $n$ , we are interested in the performance of  $\hat{\mathbf{S}}_n$  and the parametric bootstrap when  $p(n) \gg n$ . Clearly, if  $p$  is fixed at some finite value, our method will be valid for some sufficiently large  $n$ ; but we are concerned with the case where  $p$  is essentially infinite. If  $\mathcal{S}$  is a fixed (in  $p$ ) finite subset, which is a reasonable assumption, we have that  $P(\mathcal{S} \subseteq \hat{\mathbf{S}}_n)$  (or, alternatively, the sensitivity) converges to 1 as the sample size converges to infinity. This is true regardless of the rate at which  $p(n)$  converges to infinity. However, the positive predictive value still may not converge to one; that is, the number of false positives may not converge to zero. It is not enough for the target subset to be merely *contained* in the sample subset; we want the two sets to be identical with probability tending to one. To achieve this convergence, we require uniform consistency of  $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$  for  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . In summary, for a typical subset rule  $\mathbf{S}(\cdot, \cdot)$  and under the assumption that there are no subset elements on the boundary, uniform consistency of  $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$  for  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  implies that  $P(\hat{\mathbf{S}}_n = \mathcal{S}) \rightarrow 1$  as  $n \rightarrow \infty$ .

In order to control the error in  $\hat{\mathbf{S}}_n$  as an estimate of  $\mathcal{S}$  one needs to control the uniform distance between  $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$  and  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . In particular, if one wants to control the probability of finding extremely false positives in  $\hat{\mathbf{S}}_n$ , then one needs  $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$  to be within a specified distance  $\epsilon$  from the true  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , where  $\epsilon$  will depend on the definition of an extremely false positive. For example, consider the simple subset rule  $\mathbf{S}(\boldsymbol{\mu}) = \{j : \mu_j > \log 3\}$ . Then one might define an extremely false positive as a gene  $j$  with  $\mu_j < \log 2$ . In this case, given a small user-supplied number  $\delta > 0$ , one wants to choose the sample size  $n$  such that the probability that the uniform distance between  $\hat{\boldsymbol{\mu}}_n$  and  $\boldsymbol{\mu}$  is smaller than  $\epsilon = \log 3 - \log 2 \approx 0.41$  is larger than  $1 - \delta$ .

The next two theorems establish the uniform consistency of  $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$  (and, therefore,  $\hat{\mathbf{S}}_n$ ) and provide a sample size formula, respectively; both proofs rely on Bernstein’s Inequality for sums of independent mean-zero random variables. Recall that (see van der Vaart and Wellner, 1996, page 102): If  $Z_1, \dots, Z_n$  are independent, have range within  $[-W, W]$ , and have zero means, then

$$P(|Z_1 + \dots + Z_n| > x) \leq 2 \exp\left(\frac{-x^2/2}{v + Wx/3}\right) \quad (2)$$

for  $v \geq \text{var}(Z_1 + \dots + Z_n)$ .

**THEOREM 3.1 (CONSISTENCY).** *Let  $p = p(n)$  be such that  $n/\log(p(n)) \rightarrow \infty$  as  $n \rightarrow \infty$  and  $M < \infty$  (recall that  $M$  bounds the absolute value of the underlying data). As  $n \rightarrow \infty$ , then*

$$\max_j |\hat{\mu}_j - \mu_j| \rightarrow 0 \text{ in probability}$$

and

$$\max_{ij} |\widehat{\Sigma}_{ij} - \Sigma_{ij}| \rightarrow 0 \text{ in probability.}$$

This implies the following: Suppose that the subset rule  $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow \mathbf{S}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  is continuous in the sense that if, for the sequence  $(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)$  ( $p(n)$  vectors and  $p(n) \times p(n)$  matrices, respectively),

$$\|(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n) - (\boldsymbol{\mu}, \boldsymbol{\Sigma})\|_{max} \rightarrow 0, \text{ then } \|\mathbf{S}(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n) - \mathbf{S}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\|_{max} \rightarrow 0. \quad (3)$$

Then for any  $\epsilon > 0$ ,

$$P(\|\widehat{\mathbf{S}}_n - \mathcal{S}\|_{max} > \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

For example, consider subset rule 1, defined in section 2, indexed by user-supplied  $C_\mu$  and  $C_\rho$ . If there exists an  $\epsilon > 0$  such that

$$\begin{aligned} \{j : \mu_j \in (C_\mu - \epsilon, C_\mu + \epsilon)\} &= \emptyset \\ \{(i, j) : \rho_{ij} \in (C_\rho - \epsilon, C_\rho + \epsilon)\} &= \emptyset, \end{aligned} \quad (*)$$

then, for such a subset rule, we have

$$P(\widehat{\mathbf{S}}_n = \mathcal{S}) \rightarrow 1 \text{ as } n \rightarrow \infty.$$

**THEOREM 3.2 (SAMPLE SIZE FORMULA).** Let  $\epsilon > 0, 1 > \delta > 0$ , and the number of genes  $p$  be given. Let  $\sigma^2$  be an upper bound of  $\max_j \sigma_j^2$  and let  $W > 0$  be a constant such that  $P(Y_j - \mu_j \in [-W, W]) = 1$ , for all  $j$ . Define  $n^*(p, \epsilon, \delta, W, \sigma^2)$  as follows:

$$n^*(p, \epsilon, \delta, W, \sigma^2) = \frac{1}{c} (\log p + \log \frac{2}{\delta}), \text{ where } c = c(\epsilon, \sigma^2, W) = \frac{\epsilon^2}{2\sigma^2 + 2W\epsilon/3}.$$

If  $n > n^*$ , then

$$P(\max_j |\widehat{\mu}_j - \mu_j| > \epsilon) < \delta.$$

Similarly, if  $n > n^*(p^2, \epsilon, \delta, W^2, \sigma_\Sigma^2)$ , where  $\sigma_\Sigma^2$  is an upper bound of the variance of  $Y_i Y_j$ , then

$$P(\max_{ij} |\widehat{\Sigma}_{ij} - \Sigma_{ij}| > \epsilon) < \delta.$$

The consistency of  $\mathbf{S}(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)$  is a direct consequence of the uniform consistency of  $(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)$ . We will prove the uniform consistency of  $\widehat{\boldsymbol{\mu}}_n$ . For a particular component of  $\widehat{\boldsymbol{\mu}}_n$ , application of Bernstein's Inequality gives:

$$P(|\widehat{\mu}_j - \mu_j| > \epsilon) \leq 2 \exp\left(\frac{-n\epsilon^2}{2\sigma^2 + 2W\epsilon/3}\right).$$

Since  $P(\cup_j A_j) \leq \sum_j P(A_j)$ , we have an upper bound on the probability that the uniform distance from  $\widehat{\boldsymbol{\mu}}_n$  to  $\boldsymbol{\mu}$  exceeds  $\epsilon > 0$ :

$$P(\max_j |\widehat{\mu}_j - \mu_j| > \epsilon) \leq \sum_j P(|\widehat{\mu}_j - \mu_j| > \epsilon) \leq 2p \exp\left(\frac{-n\epsilon^2}{2\sigma^2 + 2W\epsilon/3}\right). \quad (4)$$

The expression on the right converges to zero if  $n/\log(p(n)) \rightarrow \infty$  as  $n \rightarrow \infty$ . A similar argument holds for  $\widehat{\Sigma}_n$ .

The  $n^*$  in theorem 3.2 is precisely the solution obtained when we set the right-hand expression of (4) equal to  $0 < \delta < 1$  and solve for  $n$ . If one assumes independence of genes, then  $P(\max_j |\widehat{\mu}_j - \mu_j| \leq \epsilon) = \prod_j P(|\widehat{\mu}_j - \mu_j| \leq \epsilon)$ . By applying Bernstein's Inequality to each term of the product, one obtains the same sample size formula as above, to first order approximation. Therefore, this sample size formula is as sharp as Bernstein's Inequality. In a similar fashion as for the mean, one obtains such a sample size formula for  $P(\max_{ij} |\widehat{E}Y_i Y_j - EY_i Y_j| > \epsilon) < \delta$  and thus for  $P(\max_{ij} |\widehat{\Sigma}_{ij} - \Sigma_{ij}| > \epsilon) < \delta$ . In this case the summation is over  $p(p-1)/2$  elements and  $Y_i Y_j$  is bounded by  $W^2$ .  $\square$

If  $p$  increases from  $p_1$  to  $p_2$ , then  $n^*(p)$  increases by a magnitude  $\log(p_2/p_1)/c$  and, if  $p$  is large, then the derivative  $\frac{d}{dp}n^*(p) = 1/(cp)$  actually converges to zero. Therefore the sample size is heavily driven by the factor  $1/c$ . Consider the example given above and suppose we want the probability of including any extremely false positives to be less than 0.1. Suppose that  $p = 5000$ , that the maximal variance is 0.5 and that the truncation level  $W$  is 1.4 (twice the maximal standard deviation). Application of theorem 3.2 says that, if  $n > n^*(5000, 0.41, 0.1, 1.4, 0.5) \approx 95$ , then  $P(\sup_j |\widehat{\mu}_j - \mu_j| > 0.41) < 0.1$ . Note that the effect of a huge increase in  $p$  on the required sample size is minor: *e.g.*  $n^*(100000, 0.41, 0.1, 1.4, 0.5) = 120$ .

To convey a general sense of the implications of this sample size formula, we provide a few examples:

$$\begin{aligned} n^*(p = 5000, \epsilon = 0.1, \delta = 0.10, M = 2, \sigma^2 = 0.5) &\approx 1304 \\ n^*(p = 5000, \epsilon = 0.5, \delta = 0.10, M = 2, \sigma^2 = 0.5) &\approx 77 \\ n^*(p = 5000, \epsilon = 0.5, \delta = 0.01, M = 2, \sigma^2 = 0.5) &\approx 92 \\ n^*(p = 5000, \epsilon = 1.0, \delta = 0.05, M = 2, \sigma^2 = 0.5) &\approx 28 \end{aligned}$$

If we set the right-hand expression of (4) equal to  $\delta$  and solve for  $\epsilon$ , we obtain a  $1 - \delta$ -uniform confidence band for  $\mu$  with radius  $\epsilon$  for each component:

$$\widehat{\mu}_j \pm \epsilon(p, n, \delta, W, \sigma^2), \text{ where} \quad \epsilon = \frac{1}{2n} \left[ \frac{2W}{3} (\log p + \log \frac{2}{\delta}) + \sqrt{\left( \left( \frac{2W}{3} (\log p + \log \frac{2}{\delta}) \right)^2 + 8n\sigma^2 (\log p + \log \frac{2}{\delta}) \right)} \right] \quad (5)$$

It is also possible to construct a confidence band by first scaling the data to have variance one (by applying the formula to  $Y_j/\sigma_j$  and using  $\sigma^2 = 1$  in (5)) and then returning to the original scale by multiplying the radius  $\epsilon$  with  $\sigma_j$  for each gene:

$$\widehat{\mu}_j \pm \sigma_j \epsilon(p, n, \delta, W, \sigma^2 = 1).$$

In the above,  $\sigma_j$  can be estimated with its empirical counterpart  $\widehat{\sigma}_{jn}$ ,  $j = 1, \dots, p$ .

Given this consistency of  $\widehat{\Sigma}_n$ , we are now concerned with the asymptotic behavior of the feature-specific probabilities as  $n \rightarrow \infty$ . Theorem 3.3 demonstrates that these probabilities converge to one (zero) when the appropriate feature is present (absent) in the target subset. For example, for genes  $j$  such that  $\mathcal{S}_j > 0$ , we have that  $p_j \rightarrow 1$ .

Table 2. Information on the data-generating parameters.

Simulation A					Simulation B				
No of genes	Mean of means	Std dev'n of means	Std dev'n	Corr within?	No of genes	Mean of means	Std dev'n of means	Std dev'n	Corr within?
30	0.9	1.2	0.6	yes	20	0.9	1.2	0.6	yes
1270	-0.7	1.4	0.7	no	20	0	2.0	0.7	yes
200	-3	2.3	1.2	no	20	-3	2.3	1.2	yes
					1440	-0.75	1.5	1	no
1500					1500				

**THEOREM 3.3.** Consider the simple subset rule 1. Let  $p = \infty$  and  $M < \infty$ . Assume that  $C_\mu$  and  $C_\rho$  are chosen so that the boundary condition (\*) of theorem 3.1 holds. Then the feature specific probabilities  $p_{j,n}$  and  $P_{ij,n}$  converge uniformly in  $i, j$  to the corresponding feature-indicators  $I(j \in \mathcal{S})$  and  $I((i, j) \in \mathcal{S})$ , as  $n \rightarrow \infty$ .

The following theorem proves consistency of the bootstrap estimates of the feature specific probabilities under the condition that  $n/\log(p(n)) \rightarrow \infty$ .

**THEOREM 3.4.** Consider the simple subset rule (1). Let  $M < \infty$ . Assume that  $C_\mu$  and  $C_\rho$  are chosen so that the boundary condition (\*) of theorem 3.1 holds. If  $n/\log(p(n))$  converges to infinity, then  $\hat{p}_{j,n}$  and  $\hat{P}_{ij,n}$  converge in probability uniformly in  $(i, j)$  to the feature-indicators  $I(j \in \mathcal{S})$  and  $I((i, j) \in \mathcal{S})$ .

In order to establish asymptotic consistency of  $(\hat{\mu}_n, \hat{\Sigma}_n)$  and validity of the bootstrap at a non-degenerate level, we have proven an infinite dimensional central limit theorem for  $\sqrt{n}((\hat{\mu}_n - \mu), (\hat{\Sigma}_n - \Sigma))$ , in which the latter are treated as elements of an infinite dimensional Hilbert space with a weighted Euclidean norm. Subsequently, we prove nonparametric asymptotic validity of the parametric bootstrap for the purpose of estimating the limiting distribution of  $\sqrt{n}(\hat{\mu}_n - \mu)$ . Similarly, this can be proved for  $\sqrt{n}(\hat{\Sigma}_n - \Sigma)$  if one assumes the multivariate normal model. These proofs can be found in van der Laan and Bryan (2000).

#### 4. SIMULATION STUDY.

We created the parameters  $(\mu, \Sigma)$  of the data-generating distribution such that there exist subsets of the  $p$  genes that are correlated with one another and uncorrelated with all other genes (specifically,  $\Sigma$  is block-diagonal). We also controlled the location and spread of the means and standard deviations for each subset. In this manner we obtain parameters  $(\mu, \Sigma)$  that exhibit the features of real-world data, in which certain genes are part of a causal mechanism and, therefore, tend to be differentially expressed and correlated with one another. Many other genes are uninvolved in the process of interest and, on average, have little to no differential expression and vary independently of other genes. Table 2 provides specific information on the subsets used to generate the parameters.

We present the results from two simulations. The first employs a simple subset rule that looks only for over-expression. The second simulation is a PAM-based rule that looks for differential expression. Since the choice of  $M$  did not matter (due to the fact that a multivariate normal distribution is already lightly tailed) we report results for  $M = \infty$ . In each simulation we draw

Table 3. *Simulation summary.*

Label	$p$	$n$	Time for 1 sample iteration	Max RAM used	$ \mathcal{S} $	avg $ \hat{\mathbf{S}} $	avg $ \tilde{\mathbf{S}} $
A	1500	65	45 mins.	91MB	30	29.34	32.07
B	1500	60	66 mins.	91MB	30	26.9	26.61

Table 4. *Subset rule summary.*

Label	Pre-screen	Mid-rule	Post-screen	Ext. false positive
A	$\mu_j > \log 2.4 \approx 0.88$	none	none	$\mu_j < \log 1.25 \approx 0.22$
B	$ \mu_j  > \log 2.7 \approx 0.99$ $ \rho_{ij}  > 0.5$	PAM, $K = 3$ $D_{ij} = 1 -  \rho_{ij} $	$D_{ij} < 0.25$ for some medoid $i$	$ \mu_j  < \log 1.25 \approx 0.22$

Table 5. *Simulation results on subset quality.*

	Simulation A		Simulation B	
	Truth	Avg Bootstrap Estimate	Truth	Avg Bootstrap Estimate
Sensitivity	0.96	0.98	0.73	0.78
Positive Predictive Value	0.98	0.90	0.82	0.79
Prop. of Ext. False Pos.	0.00	0.00	0.00	0.00
Any Ext. False Pos.	0.00	0.00	0.00	0.00
Expected Lgst. Abs. Dev. (mean)	0.34	0.34	0.43	0.42

100 first-generation samples and for each sample we draw 100 bootstrap samples. The scale of these simulations is summarized in table 3. We summarize the rules used in the simulation in table 4.

In simulation B (PAM-based rule), we required sufficiently small dissimilarity with a medoid or any previously included gene. Also, we applied the rule to the truth  $(\mu, \Sigma)$  for a range of cluster values  $k$  and chose a value of 3, based on achieving the maximum average silhouette width of 0.30. Incidentally, the finding of three clusters was consistent with the three subsets of genes generated when we selected the parameters  $(\mu, \Sigma)$ . This number of clusters was enforced throughout the simulation.

The simulation results are summarized in table 5 and figure 1. We see that bootstrap estimates of single-gene probabilities and various quality measures provide good estimates of the true values. The added complexity of the subset rule in simulation B results in a more difficult estimation problem; this is readily seen in the bootstrap results. In simulation B, as opposed to simulation A, there are genes which are excluded from  $\mathcal{S}$  but still appear in the estimated subsets fairly often; this is consistent with the positive predictive value of 0.98 for simulation A versus 0.82 for simulation B. Lastly, the scatterplot in 1 demonstrates the bias of bootstrap estimates of  $p_j$  discussed in section 2.2.3, i.e. negative bias for  $p_j$  near 1 and positive bias for  $p_j$  near 0.

## 5. DATA ANALYSIS IN HUMAN ACUTE LEUKEMIA

Golub et al. (1999) analyze gene expression data in human acute leukemias to demonstrate

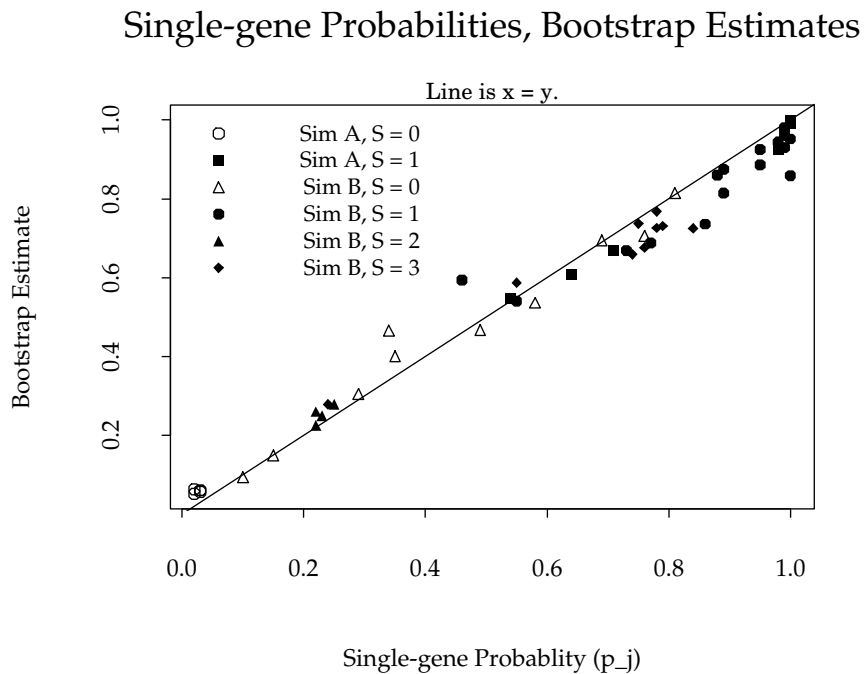
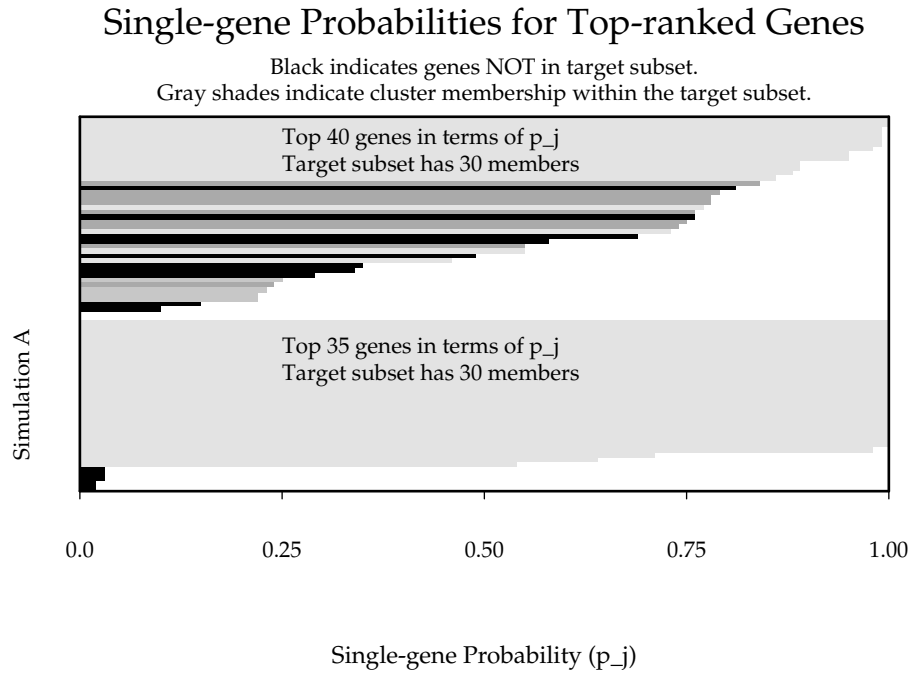


Fig. 1. Single-gene probabilities and estimates from simulation study.

a proposed method for discovering cancer classes (within a broader cancer diagnosis such as leukemia) and for predicting the class membership of a new tumor. The primary data consists of profiles for 38 leukemia patients, 27 of which have acute lymphoblastic leukemia (ALL) and 11 of which have acute myeloid leukemia (AML). For each patient there is a gene expression profile obtained from hybridization of bone marrow RNA to Affymetrix oligonucleotide microarrays. With oligonucleotide arrays, a specific probe (DNA fragment) is deposited on each spot on the array in a fixed quantity. With the cDNA arrays described earlier, there is much less control over the amount of probe placed on the array and that is the main reason for hybridizing two samples at once. By competitive hybridization, we can measure *relative* expression and avoid relying on the absolute intensity measured from one sample alone. This technical distinction means that one can actually interpret the absolute intensities from an Affymetrix chip and compare them from one patient to another.

A main goal of Golub et al. (1999) is to find a subset of genes that is effective at distinguishing between ALL and AML and that is substantially smaller than the set of genes (several thousand) placed on the array. Therefore, in this context we use our methodology to search for the subset of genes that are the best classifiers for diagnosis. It is clinically important and, apparently, difficult to distinguish the two tumor classes. Obviously, we want to look for genes which are differentially expressed in ALL patients versus AML patients. Since there is no natural pairing of measurements, we have chosen to form a reference AML expression for each gene by taking the geometric mean of the intensities across all 11 subjects. This treatment of the data emulates a typical experimental design with cDNA microarrays, in which a common reference sample is used for all experimental units. We use this as the denominator and form a ratio for each gene for all 27 ALL patients. Therefore  $n = 27$  and, after data pre-processing recommended by Golub et al. (1999), we have  $p = 5925$  genes.

We retained genes with at least 3-fold differential expression, which translates into an absolute log-ratio mean of at least 1.585. Of the original 5925 genes, 147 passed this pre-screen. We then ran PAM for several cluster numbers. The distance  $D_{ij}$  between genes  $i$  and  $j$  was defined as one minus the modified correlation proposed by Eisen et al. (1998). This quantity is obtained when one uses the normal formula for correlation  $\rho_{ij} = \sigma_{ij}/\sigma_i\sigma_j$  but replaces the means  $\mu_i$  and  $\mu_j$  with a user-specified reference value (in this case zero) in the usual calculation of covariance and standard deviation (e.g.  $\sigma'_{ij} = E(Y_i Y_j)$ ,  $\sigma'_j = \sqrt{E(Y_j^2)}$ , and  $\rho'_{ij} = \sigma'_{ij}/\sigma'_i\sigma'_j$ ). As recommended by Kaufman and Rousseeuw we chose the number of clusters  $K$  by inspecting the average silhouette widths for various values of  $K$ . For  $K = 2, 3$ , and  $4$ , the average silhouette widths were 0.87, 0.74, and 0.24 respectively (for  $K = 5, \dots, 9$ , average silhouette widths were consistently below 0.24). We decided to run the bootstrap for both  $K = 2$  and  $K = 3$  and omitted the post-screen in both cases. Genes with absolute mean less than 0.07, which corresponds to 1.05-fold differential expression or less, were deemed particularly unsuitable as classifiers and 1415 genes met this criterion in the observed data. We carried out 100 bootstrap iterations and, once the medoids for the observed data were found, the cluster centers were fixed at these medoids throughout the bootstrap.

Table 6 provides basic quality measures for the bootstrap subsets. We see that sensitivity and positive predictive value are high (around 85%) for both bootstraps and we see no extremely false positives. Figure 2 presents the single-gene proportions from the bootstrap in both the  $K = 2$  and  $K = 3$  case; the length of each horizontal bar represents the number of bootstrap iterations in which a particular gene appears in the bootstrap subset. Since the cluster centers were fixed, we can also report the stability of cluster labels and the relative frequency of each label is depicted by the shading within the horizontal bars. We see that, when increasing the cluster number from 2 to 3, in fact we split one existing cluster into two and leave one cluster untouched. In both

Table 6. *Data analysis bootstrap results on subset quality.*

	Avg Bootstrap Estimate	
	$K = 2$	$K = 3$
Sensitivity	0.88	0.88
Positive Predictive Value	0.85	0.84
Prop. of Ext. False Pos.	0.00	0.00
Any Ext. False Pos.	0.00	0.00
0.90 quantile of max abs dev. (mean)	1.20	1.25
0.90 quantile of max abs dev. (std dev'n)	3.34	3.31
0.90 quantile of max abs dev. (corr)	1.00	1.00
0.90 quantile of max abs dev. (covar)	11.15	10.93

Table 7. *Data analysis bootstrap results on cluster stability.*

Cluster	In $\widehat{S}$		$K = 2$ Bootstrap Avg.			In $\widehat{S}$		$K = 3$ Bootstrap Avg.		
	Medoid	Size	Size	Sens.	Pred. Value	Medoid	Size	Size	Sens.	Pred. Value
1	1936	105	106.2	0.90	0.89	1936	105	106.3	0.89	0.89
2	5706	42	47.6	0.85	0.78	2227	12	13.9	0.83	0.76
3						3816	30	34.2	0.81	0.75
		147	153.8	0.88	0.85		147	154.3	0.88	0.84

cases, the genes in the estimated subset reappear extremely often and almost always carry the same label as in the estimated subset. Overall, the stability of these clusters is quite strong. This is confirmed by the cluster-specific quality measures presented in table 7.

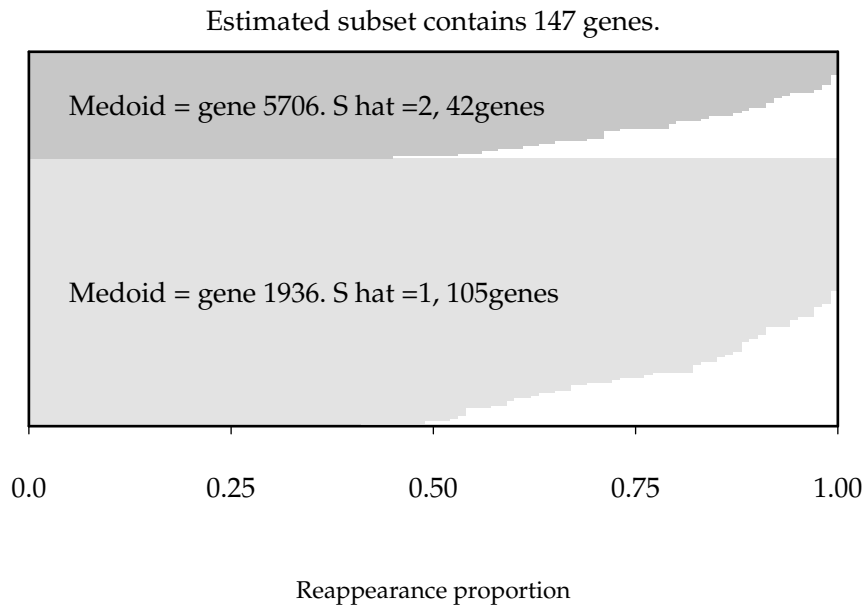
Finally, it is important to test the performance of our method at choosing genes which are effective classifiers for the ALL/AML distinction. When we limit the data available on each subject to the 50 genes that reappeared most often in the bootstrap subsets, the clustering algorithm PAM heavily favors a two-cluster structure ( $K = 2$ ) and the two clusters found correspond exactly to the ALL (27 subjects) and AML (11 subjects) groups based on actual clinical diagnosis.

Another treatment of this data that we find interesting is to define the subset rule as a function of the parameters of the two gene expression distributions produced by the ALL and AML subpopulations,  $(\mu_{ALL}, \Sigma_{ALL})$  and  $(\mu_{AML}, \Sigma_{AML})$ . In the parametric bootstrap, we would re-sample from both underlying population distributions. Relevant pre-screens would select genes with different expression distributions in the two groups; a t-test for equality of means is one example. In the mid-rule, one might choose to run clustering algorithms on each group separately and on pooled data. The post-screens could be based on the clustering and dissimilarity information both within and across groups. In fact, experiments that compare different groups and in which there is no inherent pairing are an important data structure in microarrays and we are extending both the methodology and software to handle such data.

## 6. DISCUSSION.

We would like to emphasize the distinction between the general framework we have presented and the specific choices we have made regarding particular parametric models and subset

### Reappearance proportions and cluster reproducibility



### Reappearance proportions and cluster reproducibility

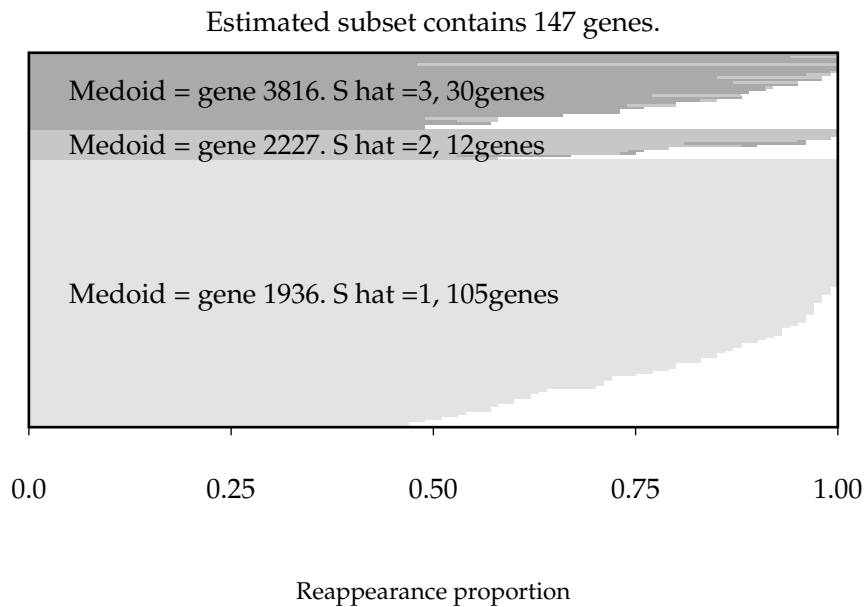


Fig. 2. Single-gene proportions and cluster stability from leukemia data analysis.

rules. By “general framework” we mean the notion of identifying those parameters of the data-generating model that are relevant (we chose the first two moments) and describing a subset of interest, and perhaps refinements within that subset, as a deterministic function of those parameters. The exact parametric assumptions (we chose multivariate normal) will then determine the approach for executing a bootstrap. We believe that many data structures arising from gene expression data can be accommodated inside this framework, although new choices of models and subset rules will require new proofs of asymptotic results like those presented here. For example, we are currently extending this framework to handle data with multiple observations per subject, such as repeated measures or dose-response data.

The field of large-scale gene expression analysis is relatively new. Naturally, this implies a great deal of exploratory data analysis based on “a holistic approach . . . that focuses on illuminating order in the entire set of observations” (Eisen et al., 1998). The application of clustering algorithms to an observed dataset, or perhaps aggregated data from several experiments, can provide extremely valuable and manageable snapshots of an entire genome (Eisen et al., 1998). This type of goal has placed great value on measuring gene expressions across as many conditions as possible and has lead, perhaps inadvertently, to a very real lack of data that includes replicates. It is important to remember that, for an organism with  $p$  genes, each hybridized chip corresponding to a fixed set of conditions contributes just one look at a  $p$ -dimensional object. In such a high-dimensional space, each additional observation gets us closer to truly understanding the interrelationships of the  $p$  genes. And so it is crucial to begin collecting data in such a way that allows the estimation of parameters such as covariance. As the techniques of gene expression analysis are applied to more refined questions, it will become increasingly important to design experiments for specific goals and to attach confidence levels to the results of data analysis.

#### REFERENCES

- Jean-Michel Claverie. Computational methods for the identifications of differential and coordinated gene expression. *Human Molecular Genetics*, 8(10):1821–1832, 1999.
- Joseph DeRisi et al. Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14:457–460, December 1996.
- Bradley Efron and Robert Tibshirani. The problem of regions. *Ann. Statist.*, 26(5):1687–1718, 1998. ISSN 0090-5364.
- Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, New York, 1993. ISBN 0-412-04231-2.
- M. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95:14863–14868, 1998.
- J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791, 1985.
- The Chipping Forecast. The chipping forecast. *Nature Genetics*, 21(1, suppl.), 1999.
- Evarist Giné and Joel Zinn. Bootstrapping general empirical measures. *Ann. Probab.*, 18(2): 851–869, 1990. ISSN 0091-1798.
- T.R. Golub, D.K. Slonim, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:321–531, October 15 1999.

- Ralf Herwig et al. Large-scale clustering of cDNA-fingerprinting data. *Genome Research*, 9:1093–1105, 1999.
- L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, 1990.
- Eliot Marshall. Do-it-yourself gene watching. *Science*, 286:444–447, 1999.
- Mark J. van der Laan and Jenny Bryan. Gene expression analysis with the parametric bootstrap. Technical Report 81, Group in Biostatistics, University of California, January 2000. Under revision for *Biostatistics*.
- Aad van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

[Received December 11, 2000. Accepted 22 January 1997. Revised 22 January 1997]