Statistics and Its Interface Volume 4 (2011) 403–415

Invisible fence methods and the identification of differentially expressed gene sets

Jiming Jiang, Thuan Nguyen and J. Sunil Rao*

The fence method (Jiang et al. 2008; *Ann. Statist.* 36, 1669–1692) is a recently developed strategy for model selection. The idea involves a procedure to isolate a subgroup of what are known as correct models (of which the optimal model is a member). This is accomplished by constructing a statistical *fence*, or barrier, to carefully eliminate incorrect models. Once the fence is constructed, the optimal model is selected from amongst those within the fence according to a criterion which can be made flexible. The construction of the fence can be made adaptively to improve finite sample performance. We extend the fence method to situations where a true model may not exist or be among the candidate models. Furthermore, another look at the fence methods leads to a new procedure, known as *invisible fence* (IF). A fast algorithm is developed for IF in the case of subtractive measure of lack-of-fit. The main focus of the current paper is microarray gene-set analysis. In particular, Efron and Tibshirani (2007; *Ann. Appl. Statist.* 1, 107–129) proposed a gene set analysis (GSA) method based on testing the significance of gene-sets. In typical situations of microarray experiments the number of genes is much larger than the number of microarrays. This special feature presents a real challenge to implementation of IF to microarray gene-set analysis. We show how to solve this problem in this paper, and carry out an extensive Monte Carlo simulation study that compares the performances of IF and GSA in identifying differentially expressed gene-sets. The results show that IF outperforms GSA, in most cases significantly, uniformly across all the cases considered. Furthermore, we demonstrate both theoretically and empirically the consistency property of IF, while pointing out the inconsistency of GSA under certain situations. An application in tracking pathway involvement in late vs earlier stage colon cancers is considered.

Keywords and phrases: Fast algorithm, Finite sample performance, Invisible fence, Limited bootstrap, Microarray gene set analysis, Model selection, Signal-consistency, Subtractive measure.

1. INTRODUCTION

The fence method (Jiang et al. 2008) is a recently developed strategy for model selection. The idea involves a procedure to isolate a subgroup of what are known as correct models (of which the optimal model is a member). This is accomplished by constructing a statistical *fence*, or barrier, to carefully eliminate incorrect models. Once the fence is constructed, the optimal model is selected from amongst those within the fence according to a criterion which can be made flexible. The construction of the fence can be made adaptively, leading to the adaptive fence method of Jiang et al. (2008). Jiang, Nguyen and Rao (2009a) simplified the (adaptive) fence procedure, in which a fence is constructed by the inequality

$$\hat{Q}(M) - \hat{Q}(\tilde{M}) \le c, \tag{1}$$

where $\hat{Q}(M) = \inf_{\theta_M \in \Theta_M} Q(M, y, \theta_M)$, Q is a measure of lack-of-fit that depends on M, a candidate model, y, the vector of observations, and θ_M, the vector of parameters under M; Θ_M is the parameter space under M; and $\tilde{M}$ is a candidate model that minimizes $\hat{Q}(M)$ among all candidate models, that is, $\tilde{M} \in \mathcal{M}$ such that $\hat{Q}(\tilde{M}) = \min_{M \in \mathcal{M}} \hat{Q}(M)$ with $\mathcal{M}$ being the space of candidate models. The constant c on the right side of (1) can be chosen as fixed, or adaptively, as mentioned. Simulation results showed that the adaptive method improves finite sample performance of fence dramatically at a computational cost (Jiang et al. 2008, Jiang, Nguyen and Rao 2009a).

A critical assumption in Jiang et al. (2008) is that there exists a true model among the candidate models. Although this assumption is necessary in establishing consistency of the fence, it limits the scope of applications. In practice, a true model simply may not exist, or exist but not among the candidate models [in this regard, George Box once said that "all models are wrong (but some are useful)"]. Furthermore, in many cases, such as the microarray gene-set analysis (see below), the definition of a "model" is, by far, not as clear as in the traditional sense. To tackle the main objective of this paper, we need to first extend the fence method.

1.1 Extension of the fence methods

In the following, we do not assume the existence of a true model among the candidates. Instead, a vector $\theta_M^* \in \Theta_M$ is called an optimal parameter vector under M with respect to the measure Q if it minimizes $\mathrm{E}\{Q(M, y, \theta_M)\}$, that is,

$$\begin{aligned} \mathrm{E}\{Q(M, y, \theta_M^*)\} &= \inf_{\theta_M \in \Theta_M} \mathrm{E}\{Q(M, y, \theta_M)\} \\ &\equiv Q(M), \end{aligned} \tag{2}$$

*Corresponding author.

where the expectation is taken with respect to the distribution of y (which does not depend on M but may be unknown). A true-Q model is a model $M \in \mathcal{M}$ such that

$$Q(M) = \inf_{M' \in \mathcal{M}} Q(M'). \tag{3}$$

Note that here the true-Q model is defined as a model that provides the best approximation, or best fit to the data, which is not necessarily a true model in the traditional sense. However, the above definitions are extensions of the traditional concepts in model selection adopted by Jiang et al. (2008), among many others. The main difference is that, in Jiang et al. (2008), a measure of lack-of-fit Q must satisfy a minimum requirement that $\mathrm{E}\{Q(M, y, \theta_M)\}$ is minimized when M is a true model and θ_M a true parameter vector under M (see examples below). With the extended definition, the minimum condition is no longer required (because it is automatically satisfied). We consider two traditional examples for illustration.

Example 1. (Negative log-likelihood) suppose that the joint distribution of y belongs to a family of parametric distributions $\{P_{M,\theta_M}, M \in \mathcal{M}, \theta_M \in \Theta_M\}$, where P_{M,θ_M} has a (joint) pdf $f_M(\cdot|\theta_M)$ with respect to a σ-finite measure μ. Consider

$$Q(M, y, \theta_M) = -\log\{f_M(y|\theta_M)\}, \tag{4}$$

the negative log-likelihood function. A model M is called a true model (in the traditional sense) if $y \sim P_{M,\theta_M}$, that is, the (joint) distribution of y is P_{M,θ_M}, for some $\theta_M \in \Theta_M$. Such a θ_M is called a true parameter vector. It turns out, by the properties of the log-likelihood function, that if M is a true model, then $\theta_M^* = \theta_M$, the true parameter vector under M. Furthermore, by the same argument, it can be shown that $Q(M)$ is minimized when M is a true model. In other words, $\mathrm{E}\{Q(M, y, \theta_M)\}$ is minimized when M is a true model and θ_M a true parameter vector under M, a property required by Jiang et al. (2008) for Q to be a measure of lack-of-fit. Therefore, any true model in the traditional sense is a true-Q model for the Q defined by (4).

Example 2. (Residual sum of squares (RSS)) consider the problem of selecting the covariates for a linear model so that $\mathrm{E}(y) = X\beta$, where X is a matrix of covariates whose columns are to be selected from a number of candidates $X_1, \ldots, X_K$, and β is a vector of regression coefficients. A candidate model M corresponds to $X_M\beta_M$, where the columns of X_M are a subset of $X_1, \ldots, X_K$, and β_M is a vector of regression coefficients of suitable dimension. Consider

$$Q(M, y, \beta_M) = |y - X_M\beta_M|^2, \tag{5}$$

which corresponds to the RSS. A model M is a true model if $\mathrm{E}(y) = X_M\beta_M$ for some β_M, which is called a true vector of regression coefficients. It can be shown that the Q defined by (5) has the property that $\mathrm{E}\{Q(M, y, \beta_M)\}$ is minimized when M is a true model and β_M a corresponding true vector of regression coefficients. It follows that if M is a true model that corresponds to X_M, then $\beta_M^* = \beta_M$, the true vector of regression coefficients corresponding to X_M; and any true model is a true-Q model for the Q defined by (5).

Given a measure of lack-of-fit, Q, for any $M \in \mathcal{M}$, let $\hat{\theta}_M$ be the minimizer of $Q(M, y, \theta_M)$ over $\theta_M \in \Theta_M$, that is, $\hat{\theta}_M \in \Theta_M$ and $Q(M, y, \hat{\theta}_M) = \inf_{\theta_M \in \Theta_M} Q(M, y, \theta_M) \equiv \hat{Q}(M)$. Let $\tilde{M} \in \mathcal{M}$ be such that $\hat{Q}(\tilde{M}) = \inf_{M \in \mathcal{M}} \hat{Q}(M)$. A model $M \in \mathcal{M}$ is in the fence if (1) holds, where c is a tuning constant which may be chosen adaptively (Jiang et al. 2008, Jiang, Nguyen and Rao 2009a). An optimal model is chosen from those within the fence so that it satisfies certain criterion of optimality. For example, one criterion of optimality that is often used is the minimum dimension criterion, where the dimension of a candidate model is typically defined as the number of free parameters under the model.

1.2 Microarray gene-set analysis

One important application area is microarray gene-set analysis. There has been interest, and studies, on the problem of identifying differentially expressed (d.e.) groups of genes, which we call gene-sets, from a set of microarray experiments (e.g., Subramanian et al. 2005). In particular, Efron and Tibshirani (2007) proposed a gene set analysis (GSA) method based on testing the significance of gene-sets. Suppose that there are N genes measured on n microarrays under two different experimental conditions, called control and treatment. The number N is usually large, say, at least a few thousands, while n is much smaller, say, a hundred or fewer. Here the interest is to assess the significance of predefined gene-sets, rather than individual genes, in terms of response to the treatment. The gene-sets are derived from different sources such as biological pathways. See Efron and Tibshirani (2007) for a nice discussion on the gene-set experiments as well as existing methods of gene set analysis. The general procedure of GSA is as follows. First compute a summary statistic for each gene, for example, the two-sample t-statistic. Then, a gene-set statistic is computed for each gene-set based on the summary statistics, for example, the average of the summary statistics for genes in the gene-set. The next step is called *restandardization*, by subtracting the genewise mean of the summary statistics from each gene-set statistic and then dividing the difference by the genewise standard deviation of the summary statistics. Results of restandardization are called gene-set scores. The p-values for each gene-set score and false discovery rates (FDR) applied to these p-values are then estimated by permutations of the class labels in the control and treatment. Depending on the FDR, a number of the gene-sets may be declared significant, or no gene-set is declared significant.

1.3 Outline of the paper

Our main objective is to apply the fence methods to microarray gene-set analysis. So far as we know, the former is a model selection strategy, while there seems to be very little modeling, if any, involved in the latter case. The extension in Subsection 1.1 has made it easier for the fence methods to adapt new environments, because the existence of a true model is no longer required. Still, it is not very clear how to bridge the connection to the current problem. The connection is made in Section 2, when we take another look at the fence methods. This leads to a new procedure, which we call *invisible fence*, or IF. Typical microarray analysis involves high dimensional data. For example, Subramanian et al. (2005) developed a collection of 522 gene pathways (gene-sets). If one anticipates a binary selection outcome (yes or no) for each gene-set, there are a total of 2^{522} different overall outcomes to choose from. Therefore, computation is a major issue in the underlying gene-set analysis. We develop a fast algorithm for IF to solve the computational problem.

In Section 3 we discuss implementation of IF to gene-set analysis. In particular, we show how to deal with a few practical challenges, especially in cases of high-dimensional data. In Section 4 we carry out an extensive simulation study that compares the performances of IF and GSA in identifying d.e. gene-sets. The results show that IF outperforms GSA, in most cases significantly, uniformly across all the cases considered. Furthermore, the simulation results reveal an interesting, previously unknown feature of GSA, that is, the method can breakdown in a certain situation, when the signals increase in an unbalanced manner. Here signals refer to the absolute values of the treatments in the gene-set experiment (see Subsection 1.2).

The surprising break-down of GSA leads to the consideration of a new type of theoretical property for a gene-set identification procedure, called *signal-consistency*. In Section 5 we formally define this concept, and prove that IF is signal-consistent. An application example on tracking pathway involvement in late stage versus earlier stage colon cancers is considered in Section 6. We conclude with a summary in Section 7.

2. INVISIBLE FENCE

2.1 Another look at the fence methods

In the adaptive fence procedure, one is supposed to run the fence (1) at a grid of c's, say, $c_1 < \cdots < c_K$, where K is fairly large (usually $\geq$100), and choose the optimal c corresponding to the highest frequency or probability of selection (see below for more detail). To be more specific, let us assume that the minimum dimension criterion is used in selecting the models within the fence. As it turns out, whatever the c, only a few models are possible results of the selection. These are the models that minimize $\hat{Q}(M)$ at different dimensions. To see this, let us assume, for simplicity, that the maximum dimension among the candidate models is 3. Let $M_j^\dagger$ be the model with dimension j such that $c_j = \hat{Q}(M_j^\dagger)$ minimizes $\hat{Q}(M)$ among all models with dimension j, $j = 0, 1, 2, 3$. Note that $c_3 < c_2 < c_1 < c_0$ (strictly speaking the $<$'s should be replaced by $\leq$'s but let us assume that the strict inequalities hold for the ease of illustration). The point is, any $c \geq c_0$ does not make a difference in terms of the final model selected by the fence, which is $M_0^\dagger$. Similarly, any $c_1 \leq c < c_0$ does not make a difference in terms of the final model selected by the fence, which is $M_1^\dagger$ [according to the fence procedure, one selects the model within the fence that has the minimum dimension and minimum $\hat{Q}(M)$ among the models within the fence that have the (same) minimum dimension]; any $c_2 \leq c < c_1$ does not make a difference in terms of the final model selected by the fence, which is $M_2^\dagger$; and any $c_3 \leq c < c_2$ does not make a difference in terms of the final model selected by the fence, which is $M_3^\dagger$ (and any $c < c_3$ leads to non-selection because no model is in the fence). In conclusion, any fence methods (adaptive or non-adaptive) will eventually select a model from one of the four candidates: $M_j^\dagger$, $j = 0, 1, 2, 3$. The question then is: Which one?

To solve this problem we use the idea of the adaptive fence by drawing bootstrap samples. The idea is to select the model that has the highest probability to best fit the empirical data when controlling the dimension of the model. To identify such a model we find, for each bootstrap sample, the best-fitting model at each dimension, that is, $M_j^{*\dagger}$, such that $\hat{Q}^*(M_j^{*\dagger})$ minimizes $\hat{Q}^*(M)$ for all models with dimension j, where $\hat{Q}^*$ represents $\hat{Q}$ computed under the bootstrap sample. We then compute the (relative) frequency (among the bootstrap samples) for different models selected, and the maximum relative frequency, say p_j^*, at each dimension j. Let $M_{j^*}^{*\dagger}$ be the model that corresponds to the maximum p_j^* (over different j's) and this is the model we select. In other words, if at a certain dimension we find a model that has the highest empirical probability to best fit the data, this is the model we select. As in Jiang et al. (2008), some extreme cases (in which the relative frequencies are always one) need to be handled differently (see Section 3).

Although the new method might look quite different from the fence, it actually uses implicitly the principle of the (adaptive) fence as explained above. For such a reason, the new method is called *invisible fence*, or IF. Once again, the idea may be interpreted as finding the model that has the best chance to best fit the empirical data when controlling the dimension of the model. This is consistent with the principle of the adaptive fence, provided that the minimum-dimension criterion is used in selecting models within the fence (note that if one does not control the dimension then the largest model always fits the best, but if one controls the dimension it is a different game).

An important special case is when there is no parameter vector θ_M under model M. In this case, we can skip $Q(M, y, \theta_M)$ and define the measure $\hat{Q}(M) = Q(M, y)$ in (1)

directly. For example, in microarray gene-set analysis suppose there are m gene-sets under consideration. Let $1, \ldots, m$ denote these gene-sets. A model M corresponds to a subset of $1, \ldots, m$ such that the corresponding gene-sets are "for real" (that is, d.e.), and the rest of the gene-sets are "not real" (that is, not d.e.). In this case, a model is completely specified by the subset, so there is no (need to introduce) θ_M. Therefore, in this case we can define the $\hat{Q}(M)$ in (1) directly without having to go through $Q(M, y, \theta_M)$. This is the case that we are dealing with in this paper.

2.2 A fast algorithm

As mentioned, computation is a major concern in high dimension problems. For instance, for the microarray gene-sets considered above, at a given dimension k ($1 \le k \le m$) one considers any subset of k gene-sets from $1, \ldots, m$. If m is large, as is typically the case, this could result in a large number of $\hat{Q}(M)$'s that have to be evaluated, not to mention that one has to consider a number of different k's, if not all possible k's.

We focus on the situation where there are a (large) number of candidate elements, such as gene-sets, so that each candidate model corresponds to a subset of the candidate elements. Let $1, \ldots, m$ denote the candidate elements. A measure $\hat{Q}$ is said to be *subtractive* if it can be expressed as

$$\hat{Q}(M) = s - \sum_{i \in M} s_i, \tag{6}$$

where s_i, $i = 1, \ldots, m$ are some nonnegative quantities computed from the data, M is a subset of $1, \ldots, m$, and s is some quantity computed from the data that does not depend on M. Typically we have $s = \sum_{i=1}^{m} s_i$, but the definition does not impose such a restriction. Also the nonnegativity constraint on the s_i's is only to ensure that $\hat{Q}(M)$ behaves like a measure of lack-of-fit, that is, larger model has smaller $\hat{Q}(M)$.

For a subtractive measure, the models that minimize $\hat{Q}(M)$ at different dimensions are found almost immediately. Let $r_1, r_2, \ldots, r_m$ be the ranking of the candidate elements in terms of decreasing s_i. Then, the model that minimizes $\hat{Q}(M)$ at dimension one is r_1; the model that minimizes $\hat{Q}(M)$ at dimension two is $\{r_1, r_2\}$; the model that minimizes $\hat{Q}(M)$ at dimension three is $\{r_1, r_2, r_3\}$, and so on. This is what we call a *fast algorithm* for IF. Although the algorithm is very simple and so is the argument, we show in the next subsection that it applies natually to gene-set analysis, but first let us point out a few other problems to which the fast algorithm, and hence the IF, are potentially applicable.

2.3 Gene-set analysis

Let s_i be the gene-set statistic, known as the gene-set score (Efron and Tibshirani 2007; also see Section 1.2), for the ith gene-set, $1 \le i \le m$. For example, s_i may be the maxmean statistic introduced by Efron and Tibshirani (2007). To compute the maxmeans one first obtains the positive and negative parts of each genewise summary statistic. Here the positive part of $x \in R$ is $x^+ = \max(x, 0)$, while the negative part is $x^- = -\min(x, 0)$. The means of the positive and negative parts are then taken for each gene-set, say, $\bar{s}_i^+$ and $\bar{s}_i^-$, respectively, for the ith gene-set, and the maxmean for the gene-set is subsequently defined as $s_i = \max(\bar{s}_i^+, \bar{s}_i^-)$. A subtractive measure for gene-set analysis is then given by (6). It is clear that all the requirements are satisfied with $s = \sum_{i=1}^{m} s_i$. Therefore, the fast algorithm applies to this case.

Efron and Tibshirani (2007) suggested using the restandardized maxmeans. Our simulation results (see Section 4), however, reveal some potential problems with the restandardization.

Remark 1. (Restandardization when the number of gene-sets is small) Suppose there are m gene-sets, each with a single gene, so that the last gene-set is d.e. and the rest are not. For simplicity, suppose that one actually observes the true means of the gene-set scores, which are $0, \ldots, 0, a$, where $a > 0$. Intuitively, one would expect an easier time to detect the d.e. gene-set when a gets larger. If one considers the gene-set scores without restandardization, the difference between the d.e. gene-set and any non-d.e. one is a, which increases with a and is not dependent on m (however, if the actual gene-set scores are $X_1, \ldots, X_{m-1}, X_m + a$, where the X_i's are independent random variables with means zero, the probability that the last gene-set score is larger than any other gene-set scores decreases as m increases). On the other hand, the restandardized gene-set scores are $-1/\sqrt{m}, \ldots, -1/\sqrt{m}, (m-1)/\sqrt{m}$. So, by restandardization a has disappeared from all the gene-set scores. In particular, the difference between the d.e. gene-set and any non-d.e. one is $\sqrt{m}$, which does not depend on a. However, the difference increases with m, the total number of gene-sets. In conclusion, when the number of gene-sets is large, restandardization is likely to improve the performance of gene-set detection; otherwise, if the number of gene-sets is small, restandardization may not be a good idea compared to using gene-set scores without restandardization.

3. GENE-SET IDENTIFICATION

Now we know the fast algorithm applies to gene-set analysis if the s_i's are chosen as the gene-set scores. To be more specific, let us assume that the maxmeans (Efron and Tibshirani 2007) are used as the gene-set scores. In their paper, Efron and Tibshirani showed that the maxmeans have the best overall performance as compared to other choices of gene-set scores (such as the means and absolute means). As mentioned, each subset of gene-sets corresponds to a model M, and the best subset of k gene-sets in the sense that it minimizes $\hat{Q}(M)$ of (6) for all models of dimension k is given by the top k gene-sets according to the gene-set scores. The

only thing that needs to be determined is: What is the optimal k?

According to Section 2 (second paragraph), the optimal k is determined by a bootstrap procedure, so the first problem that arises is how to bootstrap in the current situation. As in Efron and Tibshirani (2007), the data matrix, X, is $N \times n$, where N is the number of genes and n the number of microarrays, or samples. The underlying assumption is that the samples from the control group are i.i.d., and so are the samples from the treatment group (but the distributions of the samples from the two groups, of course, may be different). Note that each sample is a $N \times 1$ vector of genewise summary statistics. Therefore, a natural approach is to bootstrap separately from the control and treatment groups using the original idea of Efron (1979). For example, suppose there are 50 samples in the control and treatment groups, respectively. We draw a random sample of size 50 with replacement from $1, \ldots, 50$, and a random sample of size 50 with replacement from $51, \ldots, 100$, and then combine the two samples. The columns of X corresponding to the combined sample constitutes the bootstrap sample. This might sound straightforward; however, there is a complication. As is usually the case in microarray analysis, the dimension N of each sample is much higher than the sample size n. If one has to bootstrap such high dimensional vectors, the quality of bootstrap (in terms of convergence of the bootstrap distribution to the true underlying distribution; see Section 5) drops. In other words, bootstrapping the full N-dimensional columns of X may not yield a good approximation to the empirical probabilities used for the determination of k (see Section 2).

3.1 Limited bootstrap

To solve this problem we use the following strategy called the *limited bootstrap*. The idea is to bootstrap a small number of gene-sets initially, and gradually increase the number of bootstrapped gene-sets if necessary. Here by bootstrapping the gene-sets it means that only the rows of X corresponding to the selected gene-sets (and the columns corresponding to the bootstrap sample) are to be used in determining k (see below). To do this we need to know (i) what gene-sets? (ii) how small? and (iii) when to increase? To give an answer to (i) we use the principle of IF (see Section 2), that is, the selection is always made among the top gene-sets. Therefore, for any given number, say, l, the gene-sets to be bootstrapped are the top l gene-sets by the ranking of the gene-set scores. The answer to (iii) is suggested by the relative frequencies of IF, which are similar to those of the adaptive fence (Jiang et al. 2008). If the highest relative frequency excluding dimension zero occurs at the highest dimension being considered, it is an indication that the number of bootstrapped gene-sets needs to increase. The answer to (ii) is, again, motivated by the relative frequencies of IF, as follows. Suppose that one needs to carry out an all-subset selection up to dimension K, in which there is 1 model at dimension zero, K at dimension one, $K(K-1)/2$ at dimension two, and so on. Note that the empirical probability or relative frequency at dimension zero is always one (because there is only one candidate). Furthermore, a maximum relative frequency (excluding dimension zero) at dimension K would be an indication for the need of (iii). Therefore, any meaningful choice without going to (iii) should correspond to a peak of the relative frequency "in the middle". The smallest K that allows such a peak in the middle is 3. Now suppose that L gene-sets are bootstrapped. Usually it is not necessary to consider all possible dimensions up to L, especially in view of (iii). (The plot of the relative frequencies against all possible dimensions is typically W-shaped with the peak in the middle occurring at a fairly low dimension.) Here we consider all possible dimensions up to $[L/2]$, where $[x]$ is the integer part of x. It follows that the minimum number of bootstrapped gene-sets that allows a peak in the middle when comparing dimensions up to $[L/2]$ is $L = 6$.

3.2 Test for no gene-set

One difficulty with IF is that, when the highest frequency occurs at dimension one, the method cannot tell whether it is one gene-set, or no gene-set (that is d.e.). This has been observed in empirical studies that in both cases can result in the frequency situation. Logically, the highest relative frequency does not constitute a "peak", if it occurs at dimension one, because the relative frequency at dimension zero is always 1 (a peak, by definition, is a relative frequency that occurs at a certain dimension, which is higher than the relative frequencies at nearby dimensions). An important rule of the fence is called *conservative principle* (Jiang et al. 2008) which, in the IF case, says that whenever there are ties in the highest frequency, one should choose the highest dimension that ties for the highest frequency. Thus, if the highest relative frequency that occurs at dimension one is 1, by the conservative principle one chooses one gene-set over no gene-set. This is, however, an exception. What if the highest relative frequency occurs at dimension one, and it is less than 1?

To solve this problem we assist IF with a test, called test for no gene-set, when the situation occurs. The null hypothesis is that no gene-set is d.e. The test statistic is the maximum of the restandardized maxmeans (Efron and Tibshirani 2007) over all the gene-sets being considered (see Section 6 for a modification in the situation where the total number of gene-sets, m, is small). The critical value of the test is obtained by permutations. For example, suppose that there are 50 microarrays in the control and treatment groups, the controls being $1, \ldots, 50$ and treatments $51, \ldots, 100$. A random sample $i_1, \ldots, i_{100}$ is drawn without replacement from $1, \ldots, 100$ (i.e., a random permutation). Then, the new data matrix obtained by rearranging the columns of X according to $i_1, \ldots, i_{100}$ is a permutation sample. It is easy to see the rationale of the proposed test. If the null hypothesis is false, then at least one of the restandardized maxmeans is

expected to be higher than the nominal level and so is the maximum of them. Note that, unlike the GSA test which is for whether each individual gene-set is significant, our test is an overall assessment (whether some gene-sets are significant, or no gene-set is significant).

3.3 Dominant factor

Another problem that arises in gene-set analysis is called dominant factor. We use an example for illustration. Suppose that two gene-sets are d.e., of which the first is to a much greater extent than the second. The first gene-set is then called a dominant factor. What happens is that the relative frequency of IF at dimension one tends to be (much) higher than that at dimension two, and hence results in underfit. In other words, in this case, the IF tends to select the dominant factor and ignore the second gene-set, even though it is known to be d.e. The dominant factor usually happens when the sample size is limited—according to the asymptotic theory of IF, it can be shown (see Jiang, Nguyen and Rao 2009b) that, as $n \to \infty$, the relative frequency at dimension two goes to one (and that at any higher dimension stays strictly less than one) in the above example, so one would not expect the relative frequency at dimension one to be (much) higher than that at dimension two, no matter how dominant the first gene-set is (also recall the conservative principle; see Subsection 3.2).

Nevertheless, our main concern is finite sample performance. Consider, again, the above example. It is observed that, once the dominant factor is removed, the second gene-set begins to emerge. Therefore, a potential remedy is to apply IF, again, to the rest of the gene-sets after the dominant factor is selected. In other words, the IF procedure is carried out in a sequential way. Here, clearly, we need a stopping rule, otherwise there is a danger for overfit. Before a new round of IF is carried out we need to know whether any of the remaining gene-sets is d.e., hence a test for no gene-set (see Subsection 3.2) is performed. Still, there is a (small) chance that the test result is significant, even if no gene-set is d.e., and this can happen at any round of IF. Therefore, theoretically, there is still a chance that the sequential procedure can go on and on, even if no gene-sets is d.e. after the initial round. However, IF has provided us other useful information to stop the process when no gene-sets are d.e. Recall Subsection 3.1. The idea is to increase the number of bootstrapped gene-sets until one finds a number, say, L, so that, when considering dimensions up to $[L/2]$, the highest frequency does not occur at dimension $[L/2]$. This indicates that the number of d.e. gene-sets is no more than $[L/2]-1$, and hence sets up an upper bound for the sequential procedure to stop. In other words, whatever one does, the total number of selected gene-sets cannot exceed $[L/2]-1$.

3.4 An IF algorithm

When everything is put together, we have the following algorithm. The numerical procedure has incorporated the fast algorithm of Subsection 2.2.

1. Rank the gene-sets by the gene-set scores.
2. Obtain the bootstrap frequencies for the top gene-sets (see Subsection 3.1).
3. If the highest bootstrap frequency, p^*, does not occur at the boundaries of the dimensions considered, stop and report the top d^* gene-sets, where d^* corresponds to the highest bootstrap frequency p^*.
4. If p^* occurs at the right boundary of the dimensions considered, increase the number of bootstrapped gene-sets by 1, and repeat step 3.
5. If p^* occurs at the left boundary (i.e., 0) of the dimensions considered, test for zero gene-set (see Subsection 3.2). If the null hypothesis is not rejected, stop, and report the current gene-sets found (or that no gene-set is d.e.).
6. If, in step 5, the null hypothesis is rejected, add the current top gene-set to the current gene-sets found; remove the current gene-sets found from the candidate gene-sets and return to step 1.

3.5 Some notes

As can be seen, the IF procedure relies on ranking of the gene-sets. A potential problem with the ranking is how to deal with ties. So far, in either the simulations or the data analysis, we have not encountered a problem with the ties. This is because we have been dealing with continuous responses, and the chance of ties is zero, at least theoretically. However, if the responses are not continuous (this may happen even with continuous responses, if some kind of rounding is used), there is certainly a chance for ties. For example, consider the sequence $x = 1, 1, 1, 3, 2$, then ordered is 1, 2, 3, 5, 4. It is seen that the three tied numbers, 1, 1, 1, receive the ranks 1, 2, 3. It might seem unfair for the second and third 1's to receive the ranks 2 and 3, but the relative frequencies are not based on a single (bootstrap) realization. So, over a (large) number of bootstrap realizations, the subset corresponding to the 1, 1, 1 is expected to eventually distinguish itself from the rest of the subsets, and therefore be selected by the IF.

4. SIMULATION STUDIES

Efron and Tibshirani (2007) carried out an empirical study, in which the authors simulated $1,000$ genes and 50 samples in each of 2 classes, control and treatment. The genes were evenly divided into 50 gene-sets, with 20 genes in each gene-set. The data matrix was originally generated independently from the $N(0,1)$ distribution, then the treatment effect was added according to one of the following five scenarios:

1. All 20 genes of gene-set 1 are 0.2 units higher in class 2.
2. The first 15 genes of gene-set 1 are 0.3 units higher in class 2.
3. The first 10 genes of gene-set 1 are 0.4 units higher in class 2.

Table 1. IF vs GSA - empirical probabilities (in %) of TP (OF, UF)

		$\rho = 0$		$\rho = 0.3$	
Scenario	Method	One-Gene-Set	Two-Gene-Set	One-Gene-Set	Two-Gene-Set
Null	IF	95 (5,0)	95 (5,0)	64 (36,0)	64 (36,0)
	GSA	59 (41,0)	59 (41,0)	52 (48,0)	52 (48,0)
1	IF	80 (6,14)	68 (1,31)	80 (16,4)	88 (4,8)
	GSA	53 (37,10)	53 (25,22)	61 (36,3)	62 (30,8)
2	IF	88 (5,7)	88 (0,12)	88 (12,0)	97 (2,1)
	GSA	67 (32,1)	65 (26,9)	65 (35,0)	66 (32,2)
3	IF	87 (5,8)	84 (0,16)	83 (14,3)	96 (2,2)
	GSA	66 (31,3)	68 (24,8)	66 (33,1)	69 (27,4)
4	IF	73 (6,21)	63 (2,35)	75 (15,10)	80 (7,13)
	GSA	64 (28,8)	57 (19,24)	66 (29,5)	63 (21,16)
5	IF	87 (6, 7)	84 (0,16)	91 (9,0)	99 (0,1)
	GSA	70 (30,0)	76 (17,7)	82 (18,0)	86 (12,2)

4. The first 5 genes of gene-set 1 are 0.6 units higher in class 2.
5. The first 10 genes of gene-set 1 are 0.4 units higher in class 2, and the second 10 genes of gene-set 1 are 0.4 units lower in class 2.

We consider the same five scenarios in our simulation study. In Efron and Tibshirani's study only the first gene-set is of potential interest. We expand their one-gene-set case to a two-gene-set case, in which we duplicate the five scenarios to the second gene-set.

Also, in Efron and Tibshirani's study the genes were simulated independently. We consider, in addition to the independent case ($\rho = 0$), a case where the genes are correlated with equal correlation coefficient $\rho = 0.3$. The correlation is generated by associating with each microarray a random effect. The genes on the same microarray are then correlated for sharing the same random effect. Let x_{ij} be the (i, j) element of the data matrix, X, where i represents the gene and j the microarray, $i = 1, \ldots, 1,000$, $j = 1, \ldots, 100$. Here $j = 1, \ldots, 50$ correspond to the controls and $j = 51, \ldots, 100$ the treatments. Then, we have

$$x_{ij} = \alpha_j + \epsilon_{ij}, \tag{7}$$

where the α_j's and ϵ_{ij}'s are independent random effects and errors that are distributed as $N(0, \rho)$ and $N(0, 1 - \rho)$, respectively. It follows that each x_{ij} is distributed as $N(0, 1)$, and $\text{cor}(x_{ij}, x_{i'j}) = \rho$, $i \neq i'$. The treatment effects are then added to the right side of (7) for $j = 51, \ldots, 100$ and genes i in the given gene-set(s), as above.

We compare the performance of IF with GSA in gene-set identification. In Efron and Tibshirani's simulation study, the authors showed that the maxmean has the best overall performance as compared with other methods, including the mean, the absolute mean, GSEA (Gene Set Enrichment Analysis; Subramanian et al. 2005) and GSEA version of the absolute mean. Therefore, our comparisons focus on the best performer of GSA, that is, the maxmean. In addition to the one-gene-set and two-gene-set cases, each with the five scenarios listed above, the simulation comparisons also include the case where no gene-set is potentially interesting, that is, no treatment effect is added to any gene-set. This is what we call the null scenario. For GSA one needs to choose the FDR as well as the number of permutation samples for the test of significance of gene-sets. For IF, on the other hand, one also needs to specify the level of significance as well as the number of permutation samples for the test for no gene-set (see Subsection 3.2). The FDR and level of significance are both chosen as $\alpha = 0.05$. The number of permutations for both GSA and IF is 200 [which is the number that Efron and Tibshirani (2007) used for their simulations].

The first comparison is on the probability of correct identification, or true-positive (TP). For IF this means that the gene-sets selected match exactly those to which the treatment effects are added, which we call true gene-sets; similarly, for GSA this means that the gene-sets that are found significant are exactly those true gene-sets. Table 1 reports the empirical probability of TP based on 100 simulation runs. For example, for the Null Scenario, One-Gene-Set case, with $\rho = 0$, the numbers mean that for 95 out of the 100 simulation runs, IF selected no (0) gene-sets; while for 59 of the 100 simulation runs, GSA found no (0) gene-sets. As another example, for Scenario 2, Two-Gene-Set case, with $\rho = 0.3$, IF selected the exact two gene-sets, to which the treatment effects are added, for 97 out of the 100 simulation runs; while GSA found the exact two gene-sets for 66 out of the 100 simulation runs. Note that these are results of same-data comparisons, that is, for each simulation run, the results for both methods are based on the same simulated data. Also reported (in the parentheses) are empirical probabilities of overfit (OF, in the sense that the identified gene-sets include all the true gene-sets plus some false discoveries) and underfit (UF, in the sense that at least one of the true gene-sets is not discovered). It appears that IF has better performance than GSA in terms of TP uniformly across all the cases and scenarios. While most of the losses

Table 2. IF vs GSA - empirical MC (s.d.), MIC (s.d.): $\rho = 0$

Scenario	Method	One-Gene-Set	Two-Gene-Set
Null	IF	0 (0), .06 (.27)	0 (0), .06 (.27)
	GSA	0 (0), .47 (.61)	0 (0), .47 (.61)
1	IF	.86 (.34), .06 (.23)	1.64 (.57), .03 (.17)
	GSA	.90 (.30), .42 (.57)	1.75 (.50), .33 (.53)
2	IF	.93 (.25), .05 (.21)	1.88 (.32), .01 (.10)
	GSA	.99 (.10), .36 (.55)	1.91 (.28), .30 (.50)
3	IF	.92 (.27), .05 (.21)	1.83 (.40), .01 (.10)
	GSA	.97 (.17), .35 (.55)	1.92 (.27), .26 (.46)
4	IF	.79 (.40), .06 (.23)	1.57 (.63), .03 (.17)
	GSA	.92 (.27), .35 (.55)	1.73 (.50), .25 (.45)
5	IF	.93 (.25), .06 (.23)	1.79 (.51), 01 (.10)
	GSA	1 (0), .33 (.53)	1.93 (.25), .19 (.41)

Table 3. IF vs GSA - empirical MC (s.d.), MIC (s.d.): $\rho = 0.3$

Scenario	Method	One-Gene-Set	Two-Gene-Set
Null	IF	0 (0), .81 (1.17)	0 (0), .81 (1.17)
	GSA	0 (0), .57 (.66)	0 (0), .57 (.66)
1	IF	.96 (.19), .22 (.57)	1.92 (.27), .06 (.23)
	GSA	.97 (.17), .41 (.55)	1.92 (.27), .38 (.54)
2	IF	1 (0), .16 (.50)	1.99 (.10), .04 (.24)
	GSA	1 (0), .38 (.54)	1.98 (.14), .35 (.56)
3	IF	.97 (.17), .20 (.60)	1.98 (.14), .04 (.19)
	GSA	.99 (.10), .35 (.50)	1.96 (.19), .31 (.51)
4	IF	.90 (.30), .26 (.73)	1.87 (.33), .15 (.38)
	GSA	.95 (.21), .32 (.48)	1.83 (.40), .30 (.48)
5	IF	1 (0), .13 (.46)	1.99 (.10), 0 (0)
	GSA	1 (0), .20 (.45)	1.98 (.14), .12 (.32)

for IF are due to UF, OF appears to be the major problem for GSA. Furthermore, both methods appear to be fairly robust against correlations between genes.

Tables 2 and 3 report another set of summaries of the simulation results. Here reported are the mean numbers (over the simulation runs) of correctly identified gene-sets (MC) and those of incorrectly identified gene-sets (MIC). The standard deviations for the mean numbers are also reported (in the parentheses; once again, note that these are the standard deviations rather than the standard errors—the latter should be the s.d. divided by $\sqrt{100} = 10$, and therefore much smaller). For example, for Scenario 2, Two-Gene-Set case, with $\rho = 0.3$, the MC for IF is 1.99 (note that the true value is 2) with a s.d. of 0.10; the MIC for IF is 0.04 (there is no true value for MIC but, ideally, it should be 0) with a s.d. of 0.24. For GSA in this case, the MC is 1.98 with a s.d. of 0.14; the MIC is 0.35 with a s.d. of 0.56. It is seen that, for $\rho = 0$, GSA has higher MC but also higher MIC compared to IF. This is consistent with the observation from Table 1 that GSA tends to overfit while IF tends to underfit. On the other hand, the MC/MIC results are mixed for $\rho = 0.3$. Once again, there appear to be little difference between the case $\rho = 0$ and $\rho = 0.3$ for GSA. As for IF, the empirical MCs and MICs are both higher for $\rho = 0.3$, in most cases; however, the change does not seem to affect the overall performance.

Our next comparison focuses on consistency properties of both methods. Traditionally, consistency in model identification (including parameter estimation and model selection) involves sample size going to infinity. Such an assumption, however, is not very realistic in gene-set analysis, because the sample size n usually is much smaller than the number of genes under consideration. Therefore, we consider a different type of consistency, called signal-consistency. A gene-set identification procedure is signal-consistent if its probability of TP goes to one as the treatment effects, or signals, increase to infinity. Of course, one may not be able to increase the signals in real-life, but the point is to see if a procedure works perfectly well in the "ideal situation", which we believe is a basic property, just like consistency in the traditional sense. To investigate signal-consistency property of IF and GSA, we expand one of the cases, namely, the two-gene-set case of Scenario 5, by increasing the treatment effects in two different ways. First, we increase the signals in a balanced manner, that is, the signals increase at the same pace for both gene-sets. Next, we let the signals increase in an unbalanced manner, so that the pace is much faster for the first gene-set than for the second one. Table 4 reports the empirical probabilities of TP based on 100 simulation runs. Here the signals are expressed in the form of (a, b, c, d), where the values a, b, c, d are added to the right side of (7) for $51 \le j \le 100$ and 1st 10 genes of gene-set one, 2nd 10 genes of gene-set one, 1st 10 genes of gene-set two, and 2nd 10 genes of gene-set two, respectively. Case 1 is taken from the bottom two rows of Table 1 (two-gene-set case), which serves as a baseline. Then we see what happens when the signals increase. In cases 1–3, where the signals increase in the balanced way, both IF and GSA seem to work perfectly well as both methods show signs of signal-consistency. However, in cases 1, 4–10, where the signals increase in the unbalanced way, the empirical probability drops, and eventually falls apart for GSA, even with increasing signals. On the other hand, IF still shines in this situation, having perfect empirical probabilities of TP.

It is interesting to know what happens to GSA in the latest situation. The problem is restandardization. Efron and Tibshirani argued that restandardization is potentially important in that it takes into account the overall distribution of the individual gene scores. Our simulation studies also confirmed that restandardization improves finite sample performance in some cases, not just for GSA but for IF as well (recall the initial ranking of the gene-sets as well as the test for no gene-set in IF are based on restandardized maxmeans). However, in the situation where the gene-set scores are dominated by, say, a single gene-set, such as the above, the restandardized gene-set scores may look very different from those based on the permutation samples. Consider, for example, an extreme case where one gene-set is so

Table 4. IF vs GSA - empirical probabilities (in %) of TP with increasing signals

Case #	Signals	$\rho = 0$ IF	$\rho = 0$ GSA	$\rho = 0.3$ IF	$\rho = 0.3$ GSA
1	(0.4,−0.4,0.4,−0.4)	84	76	99	86
2	(0.5,−0.5,0.5,−0.5)	100	88	100	97
3	(1.0,−1.0,1.0,−1.0)	100	100	100	100
4	(1.0,−1.0,0.5,−0.5)	100	97	100	99
5	(1.5,−1.5,0.5,−0.5)	100	88	100	88
6	(2.0,−2.0,0.5,−0.5)	100	64	100	56
7	(2.5,−2.5,0.5,−0.5)	100	26	100	23
8	(3.0,−3.0,0.5,−0.5)	100	10	100	3
9	(3.5,−3.5,0.5,−0.5)	100	2	100	0
10	(4.0,−4.0,0.5,−0.5)	100	0	100	0

dominant that all but one gene-set score is below the overall mean used in the restandardization. It follows that all but one restandardized gene-set score is negative. On the other hand, the critical value for any FDR that is commonly in use is expected to be, at least, nonnegative. Therefore, a test based on comparing the restandardized gene-set scores with the critical value is expected to reject nothing but the null hypothesis corresponding to the dominant gene-set, and ignore the potential interest of any others (even though some of them are d.e.).

In introducing the GSA method, Efron and Tibshirani (2007) considered a situation where the same treatment effect is added to all the gene-sets. In other words, all the gene-sets are equally d.e. The authors used this example to make the point for the need of restandardization. The claim is that, in this case, there is "nothing special about any one gene-set". While the claim is arguable from a practical point of view, it would be interesting to see how the two methods, IF and GSA, work in a situation like this. Thus, as a final comparison, we simulated data according to Scenario 1 above, except that the 0.2 units are added to all the gene-sets. If, as the latest authors claimed, there is nothing special about any gene-set, one expects a procedure to identify no (zero) gene-set in this case. According to the results based on 100 simulation runs, when $\rho = 0$, the empirical probabilities of identifying zero gene-set is 50% for IF and 47% for GSA; when $\rho = 0.3$, the corresponding empirical probabilities are 58% for IF and 54% for GSA. So, in the latest comparison, the two methods performed similarly with IF doing slightly better.

Finally, regarding computational efficiency of the methods compared to each other, it takes, for example, 4.0 seconds to run the IF for a single simulation under Scenario 2, One-Gene-Set case with $\rho = 0$ (see, for example, Table 1), as compared to 3.0 seconds to run the GSA for the same simulation, on a server computer [Intel(R) Core(TM)2 Extreme CPU X9650 @ 3.00GHz]. It should be noted that our simulation codes for IF are not written by a professional programmer. Nevertheless, in terms of the computational efficiency, IF is, at least, comparable to GSA.

5. SIGNAL-CONSISTENCY

In this section we study the signal-consistency property of IF. Traditionally, consistency is defined under the assumption that the sample size goes to infinity. However, as mentioned (see the discussion in Section 4), such an assumption is not very practical for gene-set analysis. Therefore, we consider the asymptotic property in terms of signal-consistency. Such a property has been suggested by the simulation results in the previous section (see Table 4), so it is now time to establish it. We proceed under the basic assumptions of Section 3. In particular, we assume that the measure $\hat{Q}$ is subtractive, hence (6) holds for some s_i's, which we call gene-set scores. More specifically, we assume that the data can be expressed as

$$y_{jl} = \begin{cases} \mu_{j1} + \epsilon_{jl}, & 1 \le l \le n_1, \\ \mu_{j2} + \epsilon_{jl}, & n_1 + 1 \le l \le n, \end{cases} \tag{8}$$

$1 \le j \le N$, where N is the total number of genes; $1 \le l \le n_1$ and $n_1+1 \le l \le n$ correspond to the control and treatment, respectively; μ_{jk}, $k = 1, 2$ are the means of the controls and treatments, respectively; and n is the sample size (i.e., the number of microarrays). Let $n_2 = n - n_1$. Furthermore, we assume that the gene-set score s_i has the expression

$$s_i = \psi_i(\delta_i, \xi_i), \tag{9}$$

$i = 1, \ldots, m$, where m is the total number of (candidate) gene-sets; δ_i is an unknown parameter; ξ_i is a vector of random variables that does not depend on δ_i; and $\psi_i(\cdot, \cdot)$ is a function. Let $M_0 = \{1 \le i \le m : \delta_i \ne 0\}$. The gene-sets in M_0 are called differentially expressed (d.e.). We use a simple example to illustrate.

Example 3. Suppose that there are m gene-sets each with a single gene (so that $N = m$). The gene-set scores are the two-sample t-statistics, that is,

$$s_i = \frac{\bar{y}_{i2} - \bar{y}_{i1}}{s_{i,\mathrm{p},y}\sqrt{n_1^{-1} + n_2^{-1}}}$$

with

$$s_{i,\mathrm{p},y}^2 = \frac{(n_1 - 1)s_{i1}^2 + (n_2 - 1)s_{i2}^2}{n - 2},$$

where $\bar{y}_{i1} = n_1^{-1}\sum_{l=1}^{n_1} y_{il}$, $\bar{y}_{i2} = n_2^{-1}\sum_{l=n_1+1}^{n} y_{il}$, $s_{i1}^2 = (n_1 - 1)^{-1}\sum_{l=1}^{n_1}(y_{il} - \bar{y}_{i1})^2$ and $s_{i2}^2 = (n_2 - 1)^{-1}\sum_{l=n_1+1}^{n}(y_{il} - \bar{y}_{i2})^2$. It follows that

$$s_i = \frac{\delta_i + \bar{\epsilon}_{i2} - \bar{\epsilon}_{i1}}{s_{i,\mathrm{p},\epsilon}(n_1^{-1} + n_2^{-1})^{1/2}},$$

which is in the form of (9) with $\delta_i = \mu_{i2} - \mu_{i1}$, $\xi_i = (\xi_{i1}, \xi_{i2})'$, where $\xi_{i1} = \bar{\epsilon}_{i2} - \bar{\epsilon}_{i1}$ and $\xi_{i2} = s_{i,\mathrm{p},\epsilon}\sqrt{n_1^{-1} + n_2^{-1}}$ and $\psi_i(u, v) = (u + v_1)/v_2$ for $v = (v_1, v_2)'$. Note that in this case ψ_i does not depend on i.

Without loss of generality, assume that all the δ_i's are nonnegative. By signal consistency we mean that, as $\Delta = \min_{i\in M_0}\delta_i \to \infty$, the probability of identifying (exactly) M_0 as the d.e. gene-sets goes to one. Note that, although the sample size n is not required to go to infinity, it is necessary that the total number of observations, that is, Nn, goes to infinity, as $\Delta\to\infty$, in order to have signal consistency. We illustrate this with a simple example.

Example 4. (A counter example) suppose that there are three gene-sets, each consists of a single gene. Let $l = 1, 2$ be the controls and $l = 3, 4$ the treatments. Let y_{il}, $i = 1, 2, 3$, $l = 1, 2, 3, 4$ be the data so that $y_{il} = \epsilon_{il}$ for all i, l except for $i = 1$ and $l = 3, 4$, and $y_{1l} = \Delta + \epsilon_{1l}$, $l = 3, 4$, where $\Delta > 0$ andϵ_{il}'s are independent and distributed as $N(0, 1)$. The gene-set scores are defined as $s_i = 0.5(y_{i3} + y_{i4}) - 0.5(y_{i1} + y_{i2})$, $i = 1, 2, 3$. It is easy to see that s_i has the expression (9) with $\delta_1 = \Delta$, $\delta_2 = \delta_3 = 0$, ξ_i equal to s_i with y replaced by ϵ, $\psi_i(u, v) = u + v$, and $M_0 = \{1\}$. First note that there is a positive probability, say, p, such that 1) all the ϵ_{il}'s are bounded in absolute value by one; 2) $\min_{l=3,4}\epsilon_{2l} > \max_{l=3,4}\epsilon_{3l}$; 3) $\max_{l=1,2}\epsilon_{2l} < \min_{l=1,2}\epsilon_{3l}$. Let $y_l = (y_{il})_{i=1,2,3}$, $l = 1, 2, 3, 4$. Then, the bootstrap procedure draws samples with replacements, y_1^*, y_2^* from $\{y_1, y_2\}$, and y_3^*, y_4^* from $\{y_3, y_4\}$. It is easy to see that, under 1)–3), no matter what the bootstrap samples, one always has $s_1^* > s_2^* \vee s_3^*$ and $s_2^* > s_3^*$, if $\Delta > 4$. Therefore, at dimension two one always selects the gene-sets 1 and 2, hence the corresponding p^* is 1. Therefore, by the conservative principle, the probability is at least p that two gene-sets will be selected (note that, here, the full dimension, which is 3, is not considered in the IF, because the corresponding p^* is always one), no matter how large Δ is. It follows that the IF procedure is not signal consistent in this case.

In the following we assume that the IF frequencies, p^*, are compared for all dimensions $1 \le d \le K$, where $K < m$ and increases with m. This is practical because, in practice, one may not know an upper bound of the number of d.e. gene-sets. By letting K increase with m it guarantees that the range of d eventually covers $d_0 = |M_0|$. The rate at which K increases with m is subject to the constraints of the following assumptions. Let $\epsilon_l = (\epsilon_{jl})_{1\le j\le N}$, $1 \le l \le n$. Let $\hat{P}$ denote the "empirical distribution" of ϵ_l, $l = 1, \ldots, n$ that puts the mass $1/n_1^{n_1}n_2^{n_2}$ on every point in $x = (x_1, \ldots, x_n) \in R^{Nn}$, where $x_l = (x_{jl})_{1\le j\le N}$, such that $x_1, \ldots, x_{n_1} \in \{\epsilon_1, \ldots, \epsilon_{n_1}\}$ and $x_{n_1+1}, \ldots, x_n \in \{\epsilon_{n_1+1}, \ldots, \epsilon_n\}$. Note that this is not necessarily an empirical distribution in the usual sense, because the ϵ_l's may not be identically distributed, or even independent.

A1. (D.e. gene-sets) $d_0 > 0$, and the probability goes to one, as $\Delta\to\infty$, that

$$\hat{\mathrm{P}}\left(\min_{i\in M_0} s_i > \max_{i\notin M_0} s_i\right) = 1. \tag{10}$$

A2. (Non-d.e. gene-sets) the probability goes to one that

$$\hat{\mathrm{P}}\left(\min_{i\in M} s_i \ge \max_{i\in M_1\setminus M} s_i\right) < 1 \tag{11}$$

for every $M \subset M_1 = \{1, \ldots, m\} \setminus M_0$ with $1 \le |M| \le K - d_0$.

The first part of *A1* states that there is at least one d.e. gene-set. To see what the second part means, let $\epsilon^{(b)}$ denote an arbitrary bootstrap sample from the sampling space above (that consists of $n_1^{n_1}n_2^{n_2}$ points in R^{Nn}). Let $y_l^{(b)} = \mu_1 + \epsilon_l^{(b)}$, $1 \le l \le n_1$, and $y_l^{(b)} = \mu_2 + \epsilon_l^{(b)}$, $n_1 + 1 \le l \le n$, where $\mu_k = (\mu_{jk})_{1\le j\le N}$, $k = 1, 2$. Let $s_i^{(b)}$ be the corresponding s_i, $1 \le i \le m$. The notation $\forall b$ indicates for all possible bootstrap samples. To be more specific, consider Example 7. Suppose that the ϵ_{il}'s are independent and distributed as $N(0, 1)$. Redefine the value of the denominator of s_i as 1 if it is equal to zero (this change has no practical impact since the probability that the denominator is zero is 0). Let $U_1 = \max_{\forall b}\max_{1\le i\le m}|\bar{\epsilon}_{i2}^{(b)} - \bar{\epsilon}_{i1}^{(b)}|$, $U_2 = (n_1^{-1} + n_2^{-1})\max_{\forall b}\max_{1\le i\le m} s_{i,\mathrm{p},\epsilon^{(b)}}$, and U_3 be the minimum <u>nonzero</u> value of $(n_1^{-1} + n_2^{-1})s_{i,\mathrm{p},\epsilon^{(b)}}$, $1 \le i \le m$, $\forall b$. For any $\rho > 0$, find $\lambda_j > 0$, $j = 1, 2, 3$ such that $\mathrm{P}(U_j \le \lambda_j, j = 1, 2, U_3 \ge \lambda_3) > 1-\rho$. Let $\Delta_\rho = \lambda_1(1+1\vee\lambda_2/1\wedge\lambda_3)$. It can be shown that $\Delta > \Delta_\rho$, $U_j \le \lambda_j$, $j = 1, 2$ and $U_3 \ge \lambda_3$ imply $s_i^{(b)} > \lambda_1/1 \wedge \lambda_3$, $i \in M_0$ and $s_i^{(b)} \le \lambda_1/1 \wedge \lambda_3$, $i \notin M_0$ (regardless whether the denominator is zero), $\forall b$. Therefore, the probability is at least $1 - \rho$ that (10) holds, if $\Delta > \Delta_\rho$.

Note that $\Delta\to\infty$ is involved in *A1* but not in *A2*, so it may be wondered what is the limiting process in *A2*. Recall the counterexample above (Example 8) showing that the size of the data, Nn, has to increase in order to have signal consistency. This is the limiting process implicitly assumed here. We illustrate *A2* with a one-sample problem for the sake of simplicity (the basic arguments for the two-sample problem are very similar). Suppose that the gene-set scores can be expressed as $s_i = \Delta 1_{(i=0)} + n^{-1}\sum_{l=1}^n \epsilon_{il}$, $0 \le i \le m$, where $\Delta > 0$ and ϵ_{il}'s are i.i.d. with a continuous distribution. Thus, $M_0 = \{0\}$ in this case. Let A denote the complement of the event in *A2*, and $A_{k,M}$ the complement of (11). Then, $A = \cup_{k=1}^{K-1}\cup_{M\subset\{1,\ldots,m\},|M|=k} A_{k,M}$. Note that

$$A_{k,M} = \left\{\min_{i\in M} s_i^{(b)} \ge \max_{i\in\{1,\ldots,m\}\setminus M} s_i^{(b)}, \quad \forall b\right\}$$

for every $M \subset \{1, \ldots, m\}$ (note that here the gene-set indexes start from 0). Let B ($B_{k,M}$) denote A ($A_{k,M}$) with $\ge$ replaced by $>$. It can be shown that $B_{k,M} = \{\min_{i\in M}\epsilon_{il} > \max_{i\in\{1,\ldots,m\}\setminus M}\epsilon_{il},\ 1 \le l \le n\}$, $\forall M \subset \{1, \ldots, m\}$. Furthermore, given $1 \le k \le K$, the sets $B_{k,M}$ for all $M \subset \{1, \ldots, m\}$ with $|M| = k$ are disjoint, and have the same

probability, which is

$$\mathrm{P}\left(\min_{1\le i\le k}\epsilon_{il} > \max_{k+1\le i\le m}\epsilon_{il},\ \ 1\le l\le n\right)$$
$$=\prod_{l=1}^{n}\mathrm{P}\left(\min_{1\le i\le k}\epsilon_{il} > \max_{k+1\le i\le m}\epsilon_{il}\right)=\{C_k^m\}^{-n},$$

where C_k^m is the binomial coefficient of choosing k items out of m items (that is, $C_k^m = m!/k!(m-k)!$). It follows that $\mathrm{P}(A) = \mathrm{P}(B) \le \sum_{k=1}^{K-1}\sum_{M\subset\{1,\ldots,m\},|M|=k}\{C_k^m\}^{-n} = \sum_{k=1}^{K-1}\{C_k^m\}^{1-n}$, which goes to zero if either m, or n, increases, and K increases sufficiently slowly.

Theorem 1. *Under assumptions A1 and A2 we have with probability tending to one as $\Delta, B \to \infty$ that (i) $p^* = 1$ for $M = M_0$; and (ii) $p^* < 1$ for all M such that $d_0 < |M| \le K$; therefore, by the conservative principle, the IF chooses M_0 as the d.e. gene-sets.*

The proof is given in the Appendix.

6. AN APPLICATION: TRACKING PATHWAY INVOLVEMENT IN LATE STAGE VERSUS EARLIER STAGE COLON CANCERS

Microarray gene expression data from four distinct colon tissue samples was collected: Duke's B, C, D and liver METS as expressed by the Astler-Coller-Duke's staging system (Cohen et al. 1997). The Duke B's in our data set were actually Duke BSurvivors comprising patients still alive from the time of initial diagnosis and represent an intermediate stage of cancer. Stage C and D tissues represent a progressive worsening of the disease as the cancer begins to spread from the innermost tissue layer of the colon wall to the middle tissue layers and to nearby lymph nodes and other parts of the body; primarily the liver or lung. The liver METS (METS) represent the most advanced stage and are defined as metastasized Duke D colon cancers where the metstatic site is the liver (the other common metastatic site is the lung).

The dataset consisted of 104 samples of which 25 were BSurvivors, 21 were Duke C's, 35 were Duke D's and 23 were liver METS all collected at the Ireland Cancer Center of Case Western Reserve University. There were a total of $N = 59,618$ probe sets (genes) interrogated on each microarray chip. This data was previously used as a basis for development of Bayesian ANOVA for Microarrays (BAM) for detecting differentially expressing (individual) genes (Ishwaran and Rao 2003, 2005). One can however ask more subtle questions of the data than simply looking for gene lists that track different kinds of differential expression patterns (Ishwaran and Rao, 2005). One question of particular interest to colon cancer biologists is to see which colon cancer specific pathways (i.e., groups of genes acting in concert with one another) seem to be at play when comparing a bad prognosis tumor (i.e., liver METS) to a relatively good prognosis tumor (BSurvivor). Understanding which pathways are differentially expressed gives a truer picture of the biological processes that might be at play in a worsening prognosis.

To this end, we identified 4 colon cancer specific pathways–namely the glycolysis metabolism pathway, the hypoxia p53 expression pathway, the TGFβ cell signaling pathway and an ingenuity pathway analysis (IPA) network pathway. These pathways were actually hinted at from a BAM analysis. Specifically, it's been noted that inhibition of glycolysis effectively kills colon cancer cells in a hypoxic environment in which the cancer cells exhibit high glycolyticactivity (Xu et al. 2005). The TGFβ signaling pathway is involved in the control of several biological processes including cell proliferation, differentiation, migration and apoptosis. It's one of the most commonly altered pathways in human cancers. The connection is as follows: TGFβ signaling downstream targets are relevant to cancer in that their activation leads to growth arrest. Therefore, TGFβ serves as a tumor suppressor in normal intestinal epithelium. Many colorectal cancers end up being resistant to TGFβ induced growth inhibition. Interestingly, during the late stages of carcinogenesis, TGFβ can act as a tumor promoter as well. High activity of the pathway is associated with advanced stages and decreased survival (Xu and Pasche 2007). The IPA network pathway represents a collection of genes brought together by ingenuity path analysis after a first filtering of genes using BAM.

The number of genes making up these pathways was 11, 12, 16 and 33 genes, respectively. These gene-sets were subjected to an invisible fence analysis with the following results. The first round of IF found the $\hat{p}^*$ for dimension 1 was 0.95, much higher than the $\hat{p}^*$ for dimensions 2 and 3 (in the range of 0.40 to 0.50 for each). This effectively ruled out the higher dimensions (greater than 2; see the end of Subsection 3.3 and also Subsection 3.4). The IPA network gene set was identified potentially as differentially expressed (corresponding to the largest gene-set score). The next step of the analysis involved the test for no gene-set. Due to the small number of gene-sets ($m = 4$ in this case), it is more appropriate to use the maxmeans without restandardization for the test, as Example 6 shows. The test result showed that the IPA network was significant at the $\alpha = 0.05$ level. Thus, the dominant IPA network gene-set was removed and the testing done again. This time, the test was insignificant at the $\alpha = 0.05$ level and so the IF selection process was complete.

7. SUMMARY AND DISCUSSION

We extend the fence method to situations where a true model may not be among the candidate models, and thus expand the scope of applications of the fence method. Furthermore, we proposed a variation of the (adaptive) fence method, known as the invisible fence (IF), and established its consistency properties in terms of increasing signals, known as signal-consistency. The latter is an appropriate,

and desirable, asymptotic property for microarray gene-set anslysis, which is the main application area for IF that we consider in this paper. We developed a fast algorithm for IF that solves a high-dimensional computational problem. We showed how to implement IF for microarray gene-set analysis. We studied the finite sample performance of IF, and showed that it outperforms, in most cases significantly, the GSA method of Efron and Tibshirani (2007), uniformly across all the cases considered. The simulation results also demonstrated the signal consistency of IF as well as the signal inconsistency of GSA in a certain situation. We apply the IF method to a real data problem of tracking pathway involvement in late versus earlier stage colon cancers.

R software to implement the gene set analysis using the invisible fence method is available by contacting the authors directly.

APPENDIX: PROOF OF THEOREM 1

(i) *A1* implies that, with probability tending to one, $\min_{i\in M_0} s_i^{(b)} > \max_{i\notin M_0} s_i^{(b)}, \forall b$, hence $\min_{i\in M_0} s_i^* > \max_{i\notin M_0} s_i^*$ for every bootstrap sample [hereafter s_i^* denotes s_i computed under a bootstrap sample]. It is easy to see that the latter is equivalent to that M_0 are the top d_0 gene-sets for every bootstrap sample, hence $p^* = 1$ for $M = M_0$, with probability $\to 1$.

(ii) Now suppose that M_0 are the top d_0 gene-sets. Note that, for any $d > d_0$ and model M of dimension d, $s_i^*, i \in M$ are the top d bootstrapped gene-set scores iff $M \supset M_0$, and $s_i^*, i \in M \setminus M_0$ are the top $d - d_0$ bootstrapped gene-set scores among s_i^*, $i \in M_1$. Suppose that there is M with $d_0 < |M| \le K$ such that the corresponding p^* is equal to one. Then, by the above argument, this implies that there is $M \subset M_1$ with $1 \le |M| \le K - d_0$ such that $\min_{i\in M} s_i^* \ge \max_{i\in M_1\setminus M} s_i^*$ for every bootstrap sample. Let us see what is the probability for this to happen. Let $b = 1, \ldots, B$ denote the bootstrap samples, and $s_{i,b}^*$ denote s_i computed under the bth bootstrap sample. Let A_M ($A_{M,b}^*$) denote the event inside $\hat{P}$ in (11) (with s_i replaced by $s_{i,b}^*$). We have

(12)
$$
\begin{aligned}
&\mathrm{E}\left[\max_{1\le k\le K-d_0}\max_{M\subset M_1,|M|=k}\left\{\frac{1}{B}\sum_{b=1}^{B}1_{A_{M,b}^*}-\hat{P}(A_M)\right\}^2\right]\\
&\quad\le \sum_{1\le k\le K-d_0}\sum_{M\subset M_1,|M|=k}\mathrm{E}\left[\left\{\frac{1}{B}\sum_{b=1}^{B}1_{A_{M,b}^*}-\hat{P}(A_M)\right\}^2\right]\\
&\quad= \sum_{1\le k\le K-d_0}\sum_{M\subset M_1,|M|=k}\mathrm{E}[\mathrm{E}\{(\cdots)^2|\epsilon\}]\\
&\quad= \frac{1}{B}\sum_{1\le k\le K-d_0}\sum_{M\subset M_1,|M|=k}\mathrm{E}\left[\hat{P}(A_M)\{1-\hat{P}(A_M)\}\right]\\
&\quad\le \frac{1}{4B}\sum_{k=1}^{K-d_0}C_k^{d_1},
\end{aligned}
$$

where $d_1 = |M_1|$. On the other hand, note that $B^{-1}\sum_{b=1}^{B}1_{A_{M,b}^*} = 1$ and $\hat{P}(A_M) < 1$ imply $|B^{-1}\sum_{b=1}^{B}1_{A_{M,b}^*} - \hat{P}(A_M)| \ge 1/n_1^{n_1}n_2^{n_2}$ (this is because $\hat{P}$ is a discrete probability with increment $1/n_1^{n_1}n_2^{n_2}$). Let $\mathcal{E}_0$ denote the event that M_0 are the top d_0 gene-sets, $\mathcal{E}_1$ the event that $\hat{\mathrm{P}}(A_M) < 1$ for all $M \subset M_1$ with $1 \le |M| \le K-d_0$, and $\mathcal{E}_2$ the event that $B^{-1}\sum_{b=1}^{B}1_{A_{M,b}^*} = 1$ for some $M \subset M_1$ with $1 \le |M| \le K - d_0$. Then, we have

(13)
$$
\begin{aligned}
&\mathrm{P}(p^* = 1 \text{ for some } M \text{ with } d_0 < |M| \le K)\\
&\quad\le \mathrm{P}(\mathcal{E}_0^c) + \mathrm{P}(\mathcal{E}_2)\\
&\quad\le \mathrm{P}(\mathcal{E}_0^c) + \mathrm{P}(\mathcal{E}_1^c) + \mathrm{P}(\mathcal{E}_1\cap\mathcal{E}_2)\\
&\quad\le \mathrm{P}(\mathcal{E}_0^c) + \mathrm{P}(\mathcal{E}_1^c)\\
&\qquad + \mathrm{P}\left(\max_{1\le k\le K-d_0}\max_{M\subset M_1,|M|=k}\left|\frac{1}{B}\sum_{b=1}^{B}1_{A_{M,b}^*}-\hat{P}(A_M)\right|\right.\\
&\qquad\qquad\left.\ge\frac{1}{n_1^{n_1}n_2^{n_2}}\right)\\
&\quad\le \mathrm{P}(\mathcal{E}_0^c) + \mathrm{P}(\mathcal{E}_1^c) + \frac{(n_1^{n_1}n_2^{n_2})^2}{4B}\sum_{k=1}^{K-d_0}C_k^{d_1},
\end{aligned}
$$

by (12). The last term on the right side of (13) is arbitrarily small by choosing B sufficiently large; the first two terms go to zero by the result of (i) and *A2*. This completes the proof.

ACKNOWLEDGEMENTS

Jiming Jiang is partially supported by NSF grant DMS-0809127. J. Sunil Rao is partially supported by NSF grant DMS-0806076. The research of all three authors is partially supported by NIH grant R01-GM085205A1.

Received 10 June 2010

REFERENCES

Cohen, A., Minsky, B. and Schilsky, R. (1997). Cancer of the colon. *Cancer: Principles and Practice of Oncology*, 5th ed. (DeVita, V. T. J., Hellman, S. and Rosenberg, S. eds), 1144–1196.

Efron, B. (1979). Bootstrap method: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann. Appl. Statist.* **1** 107–129. MR2393843

Hwang, J. T. G. and Nettleton, D. (2003). Principal components regression with data-chosen components and related methods. *Technometrics* **45** 70–79. MR1956192

Ishwaran, H. and Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Amer. Statist. Assoc.* **98** 438–455. MR1995720

Ishwaran, H. and Rao, J. S. (2005). Spike and slab gene selection for multigroup microarray data. *J. Amer. Statist. Assoc.* **100** 764–780. MR2201009

Jiang, J., Rao, J. S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. *Ann. Statist.* **36** 1669–1692. MR2435452

Jiang, J., Nguyen, T. and Rao, J. S. (2009a). A simplified adaptive fence procedure. *Statist. Probab. Letters* **79** 625–629. MR2662277

Jiang, J., Nguyen, T. and Rao, J. S. (2009b). On consistency of the invisible fence when $n \to \infty$. Tech. Report, Dept. of Statist., Univ. of Calif., Davis, CA.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London. MR0727836

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. P., Lander, E. S. and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550.

Xing, Y., Stoilov, P., Kapur, K., Han, A., Jiang, H., Shen, S., Black, D. L. and Wong, W. H. (2008). MADS: A new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. *RNA* **14** 1–10.

Xu, R. H., Pelicano, H., Zhou, Y., Carew, J. S., Feng, L., Bhalla, K. N., Keating, M. J. and Huang, P. (2005). Inhibition of glycolysis in cancer cells:a novel strategy to overcome drug resistance associated with mitochondrial respiratory defect and hypoxia. *Cancer Res.* **65** 613–621.

Xu, Y. and Pasche, B. (2007). TGFβ signaling alterations and susceptibility to colorectal cancer. *Human Molecular Genetics* **16** 14–20.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Computational Graphical Statist.* **15** 265–286. MR2252527

Jiming Jiang
University of California
Davis
USA

Thuan Nguyen
Oregon Health and Science University
USA

J. Sunil Rao
University of Miami
USA
E-mail address: rao.jsunil@gmail.com