The false discovery rate: a variable selection perspective

Debashis Ghosh[*] Wei Chen[†]

Trivellore E. Raghuanthan[‡]

[*]University of Michigan, ghoshd@psu.edu

[†]University of Michigan Biostatistics, lisachen@umich.edu

[‡]University of Michigan, teraghu@umich.edu

The false discovery rate: a variable selection perspective

Debashis Ghosh, Wei Chen, and Trivellore E. Raghuanthan

Abstract

In many scientific and medical settings, large-scale experiments are generating large quantities of data that lead to inferential problems involving multiple hypotheses. This has led to recent tremendous interest in statistical methods regarding the false discovery rate (FDR). Several authors have studied the properties involving FDR in a univariate mixture model setting. In this article, we turn the problem on its side; in this manuscript, we show that FDR is a by-product of Bayesian analysis of variable selection problem for a hierarchical linear regression model. This equivalence gives many Bayesian insights as to why FDR is a natural quantity to consider. In addition, we relate the risk properties of FDR-controlling procedures to those from variable selection procedures from a decision theoretic framework different from that considered by other authors.

The false discovery rate: a variable selection perspective

Debashis Ghosh, Wei Chen and Trivellore Raghunathan

Department of Biostatistics, University of Michigan

1420 Washington Heights

Ann Arbor, MI 48109-2029

Abstract

In many scientific and medical settings, large-scale experiments are generating large quantities of data that lead to inferential problems involving multiple hypotheses. This has led to recent tremendous interest in statistical methods regarding the false discovery rate (FDR). Several authors have studied the properties involving FDR in a univariate mixture model setting. In this article, we turn the problem on its side; in this manuscript, we show that FDR is a by-product of Bayesian analysis of variable selection problem for a hierarchical linear regression model. This equivalence gives many Bayesian insights as to why FDR is a natural quantity to consider. In addition, we relate the risk properties of FDR-controlling procedures to those from variable selection procedures from a decision theoretic framework different from that considered by other authors.

Keywords: Gene Expression, Hypothesis Testing, Model Selection, Multiple Comparisons, Risk, Simultaneous Inference.

1. Introduction

Recently, scientific developments in areas such as genomics and brain imaging have led to experiments in which thousands of hypotheses are simultaneously tested. An example of this are DNA microarrays (Schena, 1999). These are biochips that assay the biochemical activities for thousands of genes simultaneously. One of the major tasks in studies involving these technologies is to find genes that are differentially expressed between two experimental conditions. The simplest example is to find genes that are up- or down-regulated in cancerous tissue relative to noncancerous tissue. Typically in these experiments, the number of genes, represented as spots on the biochip, is much larger than the number of independent samples in the study. Consequently, assessing differential expression in this setting involves performing several thousand hypothesis tests, which leads to the problem of multiple comparisons.

Historically, in problems involving simultaneous inference, the goal has been to control the familywise error rate (FWER) (Westfall and Young, 1993). However, in the current settings, such control is too stringent. Recently, several authors have advocated use of the false discovery rate (FDR) for the problem of testing multiple hypotheses simultaneously (Benjamini and Hochberg, 1995; Efron et al., 2001, Storey, 2002,2003; Genovese and Wasserman, 2002, Storey, Siegmund and Taylor, 2004). This quantity is different from FWER and generally leads to greater power for detecting alternative hypotheses.

In this paper, we turn the simultaneous inference problem on its side and study the link between the false discovery rate (FDR) with variable selection. We do this by using a Bayesian framework. This allows for a new motivation for the false discovery rate and connections with the literature on model selection. This also allows for consideration of FDR-controlling procedures from a decision theoretic point of view different from that considered by Storey (2003) and Genovese and Wasserman (2002). While the work of Abramovich et al. (2004) addresses related topics, our motivation is based on a Bayesian analysis of a hierarchical model, while theirs uses minimaxity ideas for a different type of model. The structure of this paper is as follows. In Section 2, a brief background on false discovery rate is given. In Section 3, we propose a hierarchical linear regression model and show that the false discovery rate falls out as a natural quantity in this model. Another hierarchical model is considered in Section 4; this leads to another characterization of the false discovery rate and links to traditional model selection criteria. In Section 5, we analyze the proposed methods from a risk analysis point of view different from that considered by other authors. We examine the finite-sample behavior of the procedures in Section 6. Finally, we conclude with some discussion in Section 7.

2. Background

Suppose we have observations $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, a random sample from $(Y, \mathbf{X})$, where $\mathbf{X}$ is a p-dimensional vector of covariates and Y is a continuous response variable. The ideas in this paper will be illustrated using this data structure. We first present a brief review of simultaneous hypothesis testing and the false discovery rate.

2.1. Multiple Testing Procedures

Suppose we are interested in testing a set of m hypotheses. Of these m hypotheses, suppose that for m_0 of them, the null is true. To guard against making too many type I errors, the familywise error rate (FWER) has typically been controlled. A review of methods for controlling this quantity can be found in Shaffer (1995). To better understand the FWER and FDR, we consider the following 2×2 contingency table:

[Note: Table 1 about here.]

Using the definitions from Table 1, the FWER is defined to be $P(V \geq 1)$, which is the probability that the number of false positives is greater than 1. The definition of FDR as put forward by Benjamini and Hochberg (1995) is

$$FDR \equiv E\left[\frac{V}{Q} \mid Q > 0\right] P(Q > 0).$$

The conditioning on the event $[Q > 0]$ is needed because the fraction V/R is not well-defined when $Q = 0$. Storey (2002) points out the problems with controlling this quantity and suggests use of the positive false discovery rate (pFDR), defined as

$$pFDR \equiv E\left[\frac{V}{Q} \mid Q > 0\right].$$

Conditional on rejecting at least one hypothesis, the pFDR is defined to be the fraction of rejected hypotheses that are in truth null hypotheses. In words, the pFDR is the rate at which discoveries are false. This quantity is analogous to type I error rates in single hypothesis testing problems.

The FDR and pFDR refer to one type of mistake that can be made during the hypothesis testing process. The other class of mistake that can be made is that while the alternative hypothesis is true, in practice we fail to reject the null hypothesis. This is similar to making a type II error. Thus, we define the false non-discovery rate (FNR) and positive false non-discovery rate (pFNR) to be

$$FNR \equiv E\left[\frac{T}{W} \mid W > 0\right] P(W > 0)$$

and

$$pFNR \equiv E\left[\frac{T}{W} \mid W > 0\right].$$

Conditional on failing to reject at least one hypothesis, the pFNR is the fraction of accepted hypotheses that are in truth alternative hypotheses. As with pFDR, we condition on $[W > 0]$ because T/W is not well-defined when $W = 0$. Most of this paper focuses on pFDR. Heuristically, pFNR can be thought of as the rate at which discoveries are missed.

Let $H_{01}, \ldots, H_{0G}$ represent the G null hypotheses to be tested, and let $p_1, \ldots, p_G$ denote the corresponding p-values. Benjamini and Hochberg (1995) propose a simple algorithm for selecting the hypotheses that are significant that controls the false discovery rate (FDR). Let α denote the rate at which it is desired to control the false discovery rate. The algorithm of Benjamini and Hochberg (1995) is then summarized in Box 1.

[Note: Box 1 about here.]

It is shown in Benjamini and Hochberg (1995) that the procedure in Box 1 controls the FDR at level α when the p-values are independent and uniformly distributed. Benjamini and Yekutieli (2001) show that the procedure in Box 1 controls the FDR at level α under more general forms of dependence. It involves replacing α by $\alpha/(\sum_{i=1}^{G} 1/i)$. Note that for large G, $\alpha/(\sum_{i=1}^{G} 1/i) \approx \alpha/\log G$.

2.2. Mixture Model Motivation and Estimation of the pFDR

Suppose we have independent test statistics $\mathbf{T} \equiv (T_1, \ldots, T_m)$ for testing m hypotheses. Define corresponding indicator variables $H_1, \ldots, H_m$ where $H_i = 0$ if the null hypothesis is true and $H_i = 1$ if the alternative hypothesis is true. We assume that $H_1, \ldots, H_m$ are a random sample from a Bernoulli distribution where for $i = 1, \ldots, m$, $P(H_i = 0) = \pi_0$. We assume that $T_i | H_i = 0 \sim f_0$ and $T_i | H_i = 1 \sim f_1$ for densities f_0 and f_1 $(i = 1, \ldots, m)$. Suppose we use the same rejection region R for testing each of the m hypotheses. By a theorem from Storey (2002), we have that

$$\begin{aligned} pFDR(R) &= P(H = 0 | T \in R) \\ &= \frac{\pi_0 P(T \in R | H = 0)}{P(T \in R)}. \end{aligned}$$

Using the same arguments, we can show that

$$\begin{aligned} pTNR(R) &= P(H = 1 | T \in R^c) \\ &= \frac{\pi_1 P(T \in R^c | H = 1)}{P(T \in R^c)}, \end{aligned}$$

where $\pi_1 = 1 - \pi_0$ and R^c is the complement of R.

REMARK 1. Treating $H_1, \ldots, H_m$ as parameters, we see that the definition of pFDR and pTNR are posterior probabilities, but they do not represent fully conditional posterior probabilities. The probability is conditional on the test statistic lying in a rejection region, which is different than fully conditioning on all the data. The latter posterior probability, $P(H = 0 | \mathbf{T})$, has been referred to as the local false discovery rate (Efron and Tibshirani, 2002). However, there is a substantial difference in interpretation between the positive false discovery rate and the local false discovery rate; the interested reader is referred to Berger and Sellke (1987) for further discussion.

REMARK 2. The framework above is what has been used by most authors to study the false discovery rate (Storey, 2002; Genovese and Wasserman, 2002; Storey et al., 2003). Genovese and Wasserman (2002) and Storey (2003) studied FDR-controlling procedures from various points of view, including a risk point of view. We will be utilizing a different framework for the derivation of our results.

REMARK 3. So far, we have assumed that $T_1, \ldots, T_m$ are independent. However, since pFDR and pTNR are probabilities, Storey (2002) and Storey et al. (2004) have shown that estimation of these quantities can be insensitive to certain forms of dependence asymptotically.

We now present a method for assessing differential expression direct estimation of the FDR using the algorithm of Storey (2002). We consider the following model:

$$E[Y_i] = \beta_{0j} + \beta_{1j} X_{ij}, \tag{1}$$

where X_{ij} is the jth ($j = 1, \ldots, p$) component of $\mathbf{X}_i$, $i = 1, \ldots, n$. Our scientific focus in (1) is making inference about β_{1g}. It is obvious that fitting (1) is equivalent to fitting univariate linear models on a gene-by-gene basis. Model (1) can be fit using ordinary least squares (OLS), yielding a set of statistics $T_{11}, \ldots, T_{1p}$, where T_{1j} is the least squares estimator of β_{1j} divided by its estimated standard error, $j = 1, \ldots, p$. If we use a normal distribution with mean 0 and variance 1 as the null distribution for testing $H_{0g} : \beta_{1g} = 0$, then we have G p-values $p_1, \ldots, p_G$. We then can apply Algorithm 1 of Storey (2002) to estimate the gene-specific FDR; it is summarized in Box 2.

[Note: Box 2 about here.]

REMARK 4. The previous authors who have addressed the behavior of false discovery rate procedures have ignored the variation in estimating the statistics $T_{11}, \ldots, T_{1p}$. In what we discuss in Sections 3 and 4, we will account for the variation in estimation using a hierarchical framework.

Based on the algorithm in Box 2, Storey et al. (2004) consider a class of FDR-controlling procedures. Define the following threshold function:

$$c_\alpha(F) = \sup\{0 \le t \le 1 : F(t) \le \alpha\},$$

where F is a function. Based on the estimate of FDR from Box 2, Storey et al. (2004) consider the thresholding rule $c_\alpha(\widehat{FDR}) \equiv \sup\{0 \le t \le 1 : \widehat{FDR}(t) \le \alpha\}$. This leads to the class of FDR controlling procedures described in Box 3.

[Note: Box 3 about here.]

Using martingale and empirical process arguments, Storey et al. (2004) demonstrate that when the p-values are independent, the thresholding rule provides strong control of the false discovery rate at level α. In addition, when the p-values satisfy an α-mixing type condition, the procedure in Box 3 provides control of the false discovery rate. When $\lambda = 0$ in their framework, one obtains the Benjamini and Hochberg (1995) procedure.

3. FDR and Variable Selection: Part I

In this section, we derive the false discovery rate from a different point of view. An alternative to fitting G models of the form (1) is to treat $\mathbf{X}_i$ as the independent variables and Y_i as the response variable for the ith subject, $i = 1, \ldots, n$. We can then consider a hierarchical normal regression model. At the first stage of the model,

$$Y_i \overset{ind}{\sim} N(\mathbf{X}_i^T \beta, \sigma^2).$$

For the second stage of the model, we introduce binary-valued latent variables $\gamma_1, \ldots, \gamma_p$; conditional on them,

$$\beta_i | \gamma_i \sim (1 - \gamma_i) N(0, \tau_i^2) + \gamma_i N(0, c_i^2 \tau_i^2),$$

where $c_1^2, \ldots, c_p^2$ and $\tau_1^2, \ldots, \tau_p^2$ are variance components. If $\gamma_j = 1$, then this indicates that that the jth covariate should be included in the model, while $\gamma_j = 0$ implies that it should be excluded from the variable. We next assume an inverse gamma (IG) conjugate prior for σ^2 and that γ_i is distributed as Bernoulli with probability p_i, $i = 1, \ldots, p$. Thus, we have the following multilevel model:

$$\begin{aligned}
Y_i &\overset{ind}{\sim} N(\mathbf{X}_i^T \beta, \sigma^2) && (2)\\
\beta_i | \gamma_i &\sim (1 - \gamma_i) N(0, \tau_i^2) + \gamma_i N(0, c_i^2 \tau_i^2) && (3)\\
\gamma_i &\overset{ind}{\sim} Be(p_i) && (4)\\
\sigma^2 &\sim IG(\nu/2, \nu/2) && (5)
\end{aligned}$$

This type of framework has been considered by George and McCulloch (1993) in their development of Bayesian variable selection procedures. Note that while model (1) is fundamentally univariate in nature, the model defined by equations (2)-(5) specifies a joint hierarchical model for (Y,$\mathbf{X}$).

Note that because we have utilized conjugate priors, the conditional distributions can be easily computed; this lends itself very easily to Gibbs sampling procedures for calculating the posterior distribution. The posterior distribution of β given $\mathbf{Y}$, σ and γ is normal with mean $\mathbf{A}_\gamma(\sigma)^{-2}\mathbf{X}^T\mathbf{X}\hat{\beta}_{LS}$ and variance $\mathbf{A}_\gamma$, where

$$\mathbf{A} = (\sigma^{-2}\mathbf{X}^T\mathbf{X} + \mathbf{D}^{-1}\mathbf{R}^{-1}\mathbf{D}^{-1})^{-1}.$$

The variance, σ^2, is sampled from its posterior given γ and β, which is inverse gamma with parameters $(n+\nu/2)$ and $\{(\mathbf{Y}-\mathbf{X}^T\beta)^T(\mathbf{Y}-\mathbf{X}^T\beta)+\nu\lambda/2\}$. Finally, the vector γ is sampled componentwise from the posterior distribution, the ith component ($i = 1,\ldots,G$) being Bernoulli with probability

$$P(\gamma_i = 1|\gamma_{(i)},\beta,\sigma) = \frac{P(\beta_i|\gamma_i=1)p_i}{P(\beta_i|\gamma_i=1)p_i + P(\beta_i|\gamma_i=0)(1-p_i)}.$$

The Gibbs sampling algorithm that cycles through these conditional distributions was proposed by George and McCulloch (1993).

From the point of view of selecting variables, we wish to consider the posterior distribution of $\gamma_1,\ldots,\gamma_p$. Based on the above model, the conditional distribution of $\hat{\beta}_l$ given $\sigma_l, \gamma_l = 0$ is normal with mean zero and variance $\sigma_l^2+\tau_l^2$, while that of $\hat{\beta}_l$ given $\sigma_l,\gamma_l = 1$ is normal with mean zero and variance $\sigma_l^2 + c_l^2\tau_l^2$. Observe that the relative heights of these two densities at zero is

$$u_l = \left\{\frac{\sigma_l^2/\tau_l^2 + c_l^2}{\sigma_l^2/\tau_l^2+1}\right\}^{1/2}.$$

It is also the case that $u_l = P(\gamma_l = 1|\hat{\beta}_l = 0)$, which is one minus the local false discovery rate (Efron and Tibshirani, 2002) of the lth variable at zero. Thus, the local FDR at zero is

$$P(\gamma_l = 0|\hat{\beta}_l = 0) \equiv 1 - u_l.$$

More generally, the false discovery rate based on $\hat{\beta}_l$ being in a critical region R is

$$FDR(R) \equiv \frac{\int_{x\in R}\{2\pi(\sigma_l^2+c_l^2\tau_l^2)\}^{-1/2}\exp\{-x^2/(\sigma_l^2+c_l^2\tau_l^2)\}dx}{\int_{x\in R}\{2\pi(\sigma_l^2+\tau_l^2)\}^{-1/2}\exp\{-x^2/(\sigma_l^2+\tau_l^2)\}dx}.$$

There are many points to note from this analysis. We have presented a characterization of the false discovery rate based on a Bayesian framework vastly different from those considered by Storey (2002) and Genovese and Wasserman (2002) and others. We have effectively turned the problem on the side by formulating a joint model for $(Y, \mathbf{X})$ instead of dealing with multiple univariate models of the form (1). Note that some type of regularization will probably required for the joint model; this is because no unique numerical solution exists for β if p is much larger than n. The Bayesian framework provides a natural method of regularization in this regard.

A second point to note is that we have utilized a variable selection framework to derive the FDR. This suggests that procedures that select variables based on controlling the FDR will have certain risk optimality properties in the hierarchical framework described above. In particular, George and Foster (1994) have developed a framework for risk analysis that will be applicable to the situation we are considering. In Section 5, we will apply results from their work to derive optimality of FDR-controlling procedures.

Third, as was mentioned in the previous section, Storey (2002) and Genovese and Wasserman (2002) considered FDR in a mixture model setting. Their model is univariate in nature, so it is not clear at all how to extend FDR to situations that are higher-dimensional. By contrast, we have formulated a joint model and have derived FDR as a univariate quantity within this joint framework. It is quite natural to extend the FDR into multiple dimensions based on the posterior distribution of γ. For example, we could consider the posterior distribution of γ_1 and γ_2 fairly easily here. It is not as clear how this extension would work in the other authors' proposals.

Note that in the framework presented here, dependence between the predictor variables is naturally incorporated into the definition of false discovery rate. As mentioned above, a Gibbs sampling algorithm can be used to derive the posterior distribution for γ. Using techniques described in Diebolt and Robert (1994) and Tierney (1994), we have the following theorem:

Theorem 1: *There exists a unique invariant distribution* $\pi(\gamma|\mathbf{y})$*,* $0 < \rho < 1$ *and* $C > 0$ *such* *that*

$$\int_G |\pi^{(m)}(\gamma|\mathbf{y}) - \pi(\gamma|\mathbf{y})| dg \le C\rho^m,$$

where m indexes the iteration of the Gibbs sampler.

A consequence of Theorem 1 is that the estimated FDR based on the output from the Gibbs sampler converges geometrically to the true FDR at the same rate as that described in the Theorem 1. The dependence structure on the covariates needed is needed to satisfy detailed balance. This includes all of the dependence structures described by Storey (2003): independence, block independence, α-mixing, etc. Because we are using a Gibbs sampling algorithm in order to derive the posterior distribution in the model, the false discovery rate can be derived fairly easily. Fixing a rejection region R, we simply count the proportion of MCMC samples in which the $\gamma = 0$ and $\beta \in R$. By Theorem 1 and the continuous mapping theorem, the estimated false discovery rate converges to the true false discovery rate.

Based on the posterior distribution described above, we can develop a univariate variable selection procedure analogous to those given in Box 1 and 3. We can rank $P(\gamma_i = 0|Y_1, \ldots, Y_n)$ $(i = 1, \ldots, G)$ and select the variables with small posterior probabilities. The algorithm is given in Box 4.

[Note: Box 4 about here.]

Note that while the ranking is based on marginal posterior probabilities (i.e, we integrate over γ_j for $j \neq i$), the dependence between the predictor variables is incorporated in the implementation of the Gibbs sampling algorithm. When the predictor variables are orthogonal, the algorithm in Box 4 is equivalent to the Benjamini and Hochberg (1995) procedure. This is because the posterior probabilities $P(\gamma_i = 0|\mathbf{Y})$ $(\mathbf{Y} = (Y_1, \ldots, Y_n))$ are monotonic functions of the absolute value of the univariate statistics from fitting (1), $i = 1, \ldots, G$. We have thus provided an alternative motivation for the Benjamini-Hochberg procedure different from that presented in Storey et al. (2004). In addition, the procedure in Box 4 will be equivalent to Benjamini-Hochberg whenever $P(\gamma_i = 0|\hat{\beta}_i \in R)$ is a monotonic function of the univariate p-values. As we will see later in Section 5, in this framework, the procedure in Box 4 will be shown to have certain optimality properties from a risk point of view.

4. FDR and Variable Selection: Part II

In this section, we formulate a slightly different hierarchical regression model in order to consider the false discovery rate. At the first stage of the model,

$$Y_i \overset{ind}{\sim} N(\mathbf{X}_i^T\beta, \sigma^2)$$

as before. We will consider σ^2 to be known here, in contrast to the model in Section 3. Again, binary-valued latent variables $\gamma_1, \ldots, \gamma_p$ are included here. The priors we consider are of the form $p(\beta_\gamma, \gamma|d, w) = p(\beta_\gamma|\gamma, d)p(\gamma|w)$, where $p(\beta_\gamma|\gamma, d)$ is the pdf of a q_γ-dimensional normal random variable with mean zero and variance $d\sigma^2(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}$ $(d > 0)$ and $p(\gamma|w)$ is the pmf of a Binomial random variable with probability w. Thus, we have the following multilevel model:

$$\begin{aligned} Y_i &\overset{ind}{\sim} N(\mathbf{X}_i^T\beta, \sigma^2) & (6)\\ \beta|\gamma, d &\sim N_{q_\gamma}(\mathbf{0}, d(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}) & (7)\\ \gamma|w &\sim Bin(g, w). & (8) \end{aligned}$$

Observe that hierarchical models (2-5) and (6-8) are different. No prior is assumed for σ^2 here, since we are treating it as known. In addition, the prior for β depends on the covariates $\mathbf{X}$. Based on (7) and (8), the parameter d controls the size of the nonzero coefficients of β, while w controls the number of coefficients that are nonzero. Smaller values of w correspond to smaller models, while larger values tend to favor less parsimonious models. This model formulation has been utilized by Smith and Kohn (1996) and George and Foster (2000). We can again perform a Bayesian analysis of (6)-(8) construct a variable selection procedure similar to that given in Box 4. The procedure is selected in Box 5.

[Note: Box 5 about here.]

In the situation where the design matrix is orthogonal, the selection procedure from Box 5 is equivalent to the Benjamini-Hochberg (1995) procedure.

As for the hierarchical model studied in the previous section, Gibbs sampling methods can be used to calculate the posterior distribution of the parameters. The posterior distribution of β given $\mathbf{Y}$, σ and γ is normal with mean $\mathbf{A}_\gamma(\sigma)^{-2}\mathbf{X}^T\mathbf{X}\hat{\beta}_{LS}$ and variance $\mathbf{A}_\gamma$, where

$$\mathbf{A} = (\sigma^{-2}\mathbf{X}^T\mathbf{X} + \mathbf{D}^{-1}\mathbf{R}^{-1}\mathbf{D}^{-1})^{-1}.$$

The variance, σ^2, is sampled from its posterior given γ and β, which is inverse gamma with parameters $(n + \nu/2)$ and $\{(\mathbf{Y} - \mathbf{X}^T\beta)^T(\mathbf{Y} - \mathbf{X}^T\beta) + \nu\lambda/2\}$. Finally, the vector γ is sampled componentwise from the posterior distribution, the ith component $(i = 1, \ldots, G)$ being Bernoulli with probability

$$P(\gamma_i = 1|\gamma_{(i)}, \beta, \sigma) = \frac{P(\beta_i|\gamma_i = 1)p_i}{P(\beta_i|\gamma_i = 1)p_i + P(\beta_i|\gamma_i = 0)(1 - p_i)}.$$

In model (6)-(8), selecting models corresponds to finding the combinations of variables with the largest posterior probabilities of γ. By the arguments of George and Foster (2000), the posterior distribution of γ, given Y, d and w is proportional to

$$\frac{d}{2(1+d)}\{SS_\gamma/\sigma^2 - H(d, w)q_\gamma\},$$

where

$$SS_\gamma = (\mathbf{Y} - \mathbf{X}_\gamma^T\beta_\gamma)^T(\mathbf{Y} - \mathbf{X}_\gamma^T\beta_\gamma)/(n - q_\gamma),$$

and $H(d, w) = d^{-1}(1 + d)[2\log\{(1 - w)/w\} + \log(1 + d)]$. Using Theorem 1 from George and Foster (2000), if $SS_{\gamma_1}/\sigma^2 - H(d, w)q_{\gamma_1} > SS_{\gamma_2}/\sigma^2 - H(d, w)q_{\gamma_2}$ for models γ_1 and γ_2, then the posterior distribution of γ_1 is larger than that of γ_2; the converse is also true. If $H(d, w) = 2$, $\log G$ or $\log n$, then selecting models based on the posterior probability is equivalent to model selection based on AIC (Akaike, 1973), BIC (Schwarz, 1978) and RIC (Foster and George, 1994).

From the previous section, we have that the local FDR for the ith variable is $P(\gamma_i = 0|\hat{\beta}_i = 0)$ and that the FDR for a given rejection region R is $P(\gamma_i = 0|\hat{\beta}_i \in R)$. Based on the hierarchical model presented here, we can motivate selection procedures based on the local FDR and FDR as model selection procedures. The univariate selection procedure described in Box 4 can be thought of as selecting between models with one independent variable. One major difference between the model selection criteria and the FDR quantities is that while the former corresponding to posterior distributions of γ given the full data, the local FDR and pFDR correspond to the posterior distribution of γ given a partial conditioning of the data. To be specific, the local FDR is the the posterior probability of γ equalling zero, given the region of the data $\mathbf{x}$ where $\hat{\beta}(\mathbf{x}) = 0$. Similarly, the pFDR is the the posterior probability of γ equalling zero, given the region of the data $\mathbf{x}$ where $\hat{\beta}(\mathbf{x})$ falls in the rejection region R. However, by the same arguments leading to Theorem 1 of George and Foster (2000), we have that ranking variables univariately based on $P(\gamma = 0|\hat{\beta} = 0)$ leads to a proper calibration. Similarly, a proper calibration is achieved by ranking variables univariately based on $P(\gamma = 0|\hat{\beta} \in R)$ for a given rejection region R.

In the model (6)-(8), we have assumed that the design matrix can allow for general dependence between the predictor variables. However, in the situation where the design matrix is orthogonal, the procedure described in Box 5 reduces to that proposed by Benjamini and Hochberg (1995).

Note that we have assumed that σ^2 is known in this discussion. In practice, this will not be the case. For this analysis, we can plug in an estimator for σ^2 in (6) that accounts for the selection procedure. Potential choices for estimators of the variance can be found in Section 1 of George and Foster (2000). In certain examples, G can be on the order of the sample size (n) or even much larger than n. An example of the former is wavelet regression (Vidakovic, 1999), while in microarray data analysis, G is much larger than n. How to estimate σ^2 in the latter setting for this model formulation remains an open question.

5. A Decision Theoretic Framework

Here, we consider the hierarchical regression model from Section 3 and study the properties of the variable selection procedure in Box 4 from a decision theoretic perspective. Much of the discussion here is based on that in Foster and George (1994). Define $R(\beta, \hat{\beta})$ to be the predictive risk of the estimator $\hat{\beta}$, i.e.

$$R(\beta, \hat{\beta}) = E_\beta |X\hat{\beta} - X\beta|^2.$$

We know that the vector γ can take 2^p possible values. Let $\zeta \equiv (\zeta_1, \ldots, \zeta_G)$ denote the true model, so $\zeta_i = I(\beta_i \neq 0)$, $i = 1, \ldots, G$. The risk inflation (Foster and George, 1994) is given by

$$RI(\gamma) \equiv \sup_{\beta} \frac{R(\beta, \hat{\beta}_\gamma)}{R(\beta, \hat{\beta}_\zeta)}. \tag{9}$$

Observe that the denominator in (9) is the lowest possible risk, since it represents the risk for the ideal model. In most variable selection settings, we first select the variables, and then estimate β using the selected variables. The risk inflation (9) reflects the worst-possible increase in risk with using a combination selection/estimation procedure. Based on this setting, we wish to find procedures that minimize (9) over a large class of procedures.

Before describing how the FDR procedures in Boxes 4 and 5 fit into this framework, we first start by considering the risk inflation for various procedures. For the sake of simplicity, we consider the case where $\mathbf{X}^T\mathbf{X}$ is diagonal. In this case, variable selection can be reduced to the situation of ranking variables based on the magnitude of the corresponding univariate statistics. Suppose we estimate β using least squares and that $n > G$. For this situation, the risk inflation is G. If we use AIC (Akaike, 1973) for variable selection, the risk inflation turns out to be approximately $0.57G$. For variable selection using BIC (Schwarz, 1978), the risk inflation is approximately $\log n$ if $G << n^{1/2}$ and $(2\log n/(\pi n))^{1/2}$ if $G >> n^{1/2}$.

Foster and George (1994) prove that for the case of diagonal $\mathbf{X}^T\mathbf{X}$, the optimal rule (i.e., the rule that minimizes (9)) is a threshold rule that selects the top $(2\log G)$ variables based on the absolute magnitude of the univariate statistics. Equivalently, the optimal threshold rule selects the $2\log G$ variables with the smallest univariate p-values. Note that $\mathbf{X}^T\mathbf{X}$ corresponds to the situation of independent statistics from §2.2. The Benjamini-Hochberg (1995) procedure is a data-dependent threshold rule that is a special case of the class of FDR-controlling procedures proposed by Storey et al. (2004) in Box 3. Thus, when $\hat{k} \approx (2\log G)$, then the Benjamini-Hochberg (1995) procedure will be optimal from a risk inflation framework.

In the general case where $\mathbf{X}^T\mathbf{X}$ is nonorthogonal, Foster and George (1994) show that the risk inflation (9) is bounded from below by $2\log G - o(\log G)$. Heuristically, we can argue that when $\hat{k} \approx 2\log G$, then the Benjamini-Yekutieli (2001) procedure will be approximately optimal in this framework as well.

6. Simulation Studies

We next sought to study the finite-sample properties of the proposed methodologies using simulation studies. We considered two situations. The first is where p is smaller than n, while the second is when p is larger than n. We considered the model from Section 2. In the first set of simulations, $n = 50$ and $p = 10$. The true model is $E(Y) = X_1 + 1.5X_2 + 3X_3$. The variance of the error term in all simulation studies is one. The predictors were generated with correlation $\rho = 0.1, 0.3, 0.5, 0.7$ and 0.9. A receiver operating characteristic (ROC) curve was constructed based on taking the top k variables ($k = 1, 2, 3, 4, 5,$ and 10) based on the estimated posterior probability from the algorithm in Box 3. The ROC curves averaged across 250 simulations for each setting are shown in Figure 1; as is shown there, ranking variables based on the estimated univariate posterior probabilities accurately identifies the true model. To study the finite-sample properties of the risk behavior of the proposed procedures, a second simulation study was done in which the true mean-squared error (MSE) was compared with estimated mean-squared errors based on selecting the top k variables. Here we considered the same model as above with $\rho = 0.5$; values of $k = 3, 5$ and 8 were considered. The MSE values are taken over 250 simulations. The results are provided in Figure 2. We find that even though a selection of $k = 3$ virtually mimics the behavior of the true MSE, selecting $k = 5$ yields on average a lower MSE. In keeping with the results of Foster and George (1994), they would suggest a model using the top $2\log(50) \approx 8$ variables. The estimated MSE from that criterion is competitive with the true MSE.

Next, the situation in which p is larger than n was considered. For this situation, we took $p = 20$ and $n = 10$. We considered the same true model as in the previous paragraph, along with the same correlation values. Cutoff values $k = 1, 2, 3, 4, 5, 10, 15$ and 20 were used. The ROC curves averaged across 250 simulations for each setting are shown in Figure 3. Based on this, we find that there is substantial difference in the performance of the procedure depending on how much correlation is in the data. More correlation leads to better performance; this is due to the fact that such a situation leads to a smaller effective dimension size of the model. Next, the mean-squared errors from variable selection procedures were considered in a manner analogous to that in Figure 2. The plot is shown in Figure 4. Because of the fact that p is bigger than n, we find that the mean-squared errors from the selection procedures are smaller than the true mean-squared error. These results suggests that procedures based on a univariate selection criterion for model selection might have nice risk properties.

7. Discussion

In this article, we have attempted to approach the false discovery rate from a different angle relative to that in the previous literature (Benjamini and Hochberg, 1995; Genovese and Wasserman, 2002; Storey, 2002). We find that the local false discovery rate is a natural quantity that arises in the variable and model selection context. By finding this link, we are then able to tie in results from the model selection literature and risk analysis. The results suggest that procedures for ranking variables for consideration in a model based on univariate posterior probability criteria behave well from a risk point of view.

Acknowledgments

The first author would like to thank Tom Nichols and John Storey for helpful discussions.

References

Abramovich, F., Benjamini, Y., Donoho, D., Johnstone, I. 2004. Adapting to unknown sparsity by controlling the false discovery rate. Technical Report, Department of Statistics, Stanford University.

Akaike, H. 1973. Information theory and an extension to the maximum likelihood principle. In *2nd International Symposium on Information Theory* (Ed. B. N. Petrov and F. Csaki), pp. 267 – 281. Budapest: Akademia Kiado.

Benjamini, Y., Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Roy. Statist. Soc. B 57, 289–300.

Benjamini, Y., Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188.

Berger, J. O., Sellke, T. 1987. Testing a point null hypothesis: the irreconcilability of p-values and evidence. J. Amer. Statist. Assoc. 82, 112 – 122.

Efron, B., Tibshirani, R., Storey, J. D., Tusher, V. 2001. Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc. 96, 1151 – 1160.

Efron, B., Tibshirani, R. 2002. Empirical Bayes methods and false discovery rates for microarrays. Genet. Epid. 23, 70 – 86.

Foster, D. P., George, E. I. 1994. The risk inflation criterion for multiple regression. Ann. Stat. 22, 1947 – 1975.

Genovese, C., Wasserman, L. 2002. Operating characteristics and extensions of the false discovery rate procedure. J. Roy. Statist. Soc. B 64, 499 – 517.

George, E. I., Foster, D. P. 2000. Calibration and empirical Bayes variable selection. Biometrika 87, 731 – 747.

George, E. I., McCulloch, R. E. 1993. Variable selection via Gibbs sampling. J. Amer. Statist. Assoc. 88, 881 – 889.

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Stat. 6, 461 – 464.

Shaffer, J. 1995. Multiple hypothesis testing. Ann. Rev. Psych. 46, 561–584.

Smith, M., Kohn, R. 1996. Nonparametric regression using Bayesian variable selection. J. Econometrics 75, 317 – 344.

Storey, J. D. 2002. A direct approach to false discovery rates. J. Roy. Statist. Soc. B 64, 479 – 498.

Storey, J. D. 2003. The positive false discovery rate: A Bayesian interpretation and the q-value. Ann. Stat. 31, 2013 – 2035.

Storey, J. D., Taylor, J. E., Siegmund, D. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J. Roy. Statist. Soc. B 66, 187 – 205.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). Ann. Stat. 22, 1701 – 1728.

Vidakovic, B. 1999. Statistical modeling by wavelets. Wiley, New York.

Table 1: Outcomes of m tests of hypotheses

	Accept	Reject	Total
True Null	U	V	m_0
True Alternative	T	S	m_1
	W	Q	m

Box 1. Benjamini and Hochberg (1995) procedure

(a) Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(G)}$ denote the ordered, observed p-values.

(b) Find $\hat{k} = \max\{1 \leq k \leq G : p_{(k)} \leq \alpha k/G\}$.

(c) If $\hat{k}$ exists, then reject null hypotheses $p_{(1)} \leq \cdots \leq p_{(\hat{k})}$. Otherwise, reject nothing.

Box 2. Proposed Algorithm for estimating pFDR and FDR

(a) Fit (1) for each gene g, $g = 1, \ldots, G$.

(b) Calculate a p-value using $\hat{\beta}_{1g}/\hat{SE}(\hat{\beta}_{1g})$, $g = 1, \ldots, G$.

(c) Let $p_1, \ldots, p_G$ denote the G p-values. Estimate π_0, the proportion of differentially expressed genes and $F_P(x)$, the cdf of the p-values by

$$\hat{\pi}_0 = \frac{W(\lambda)}{(1-\lambda)G}$$

and

$$\hat{F}_P(x) = \frac{\min\{R(\gamma), 1\}}{G},$$

where $R(\gamma) = \#\{p_i \leq \gamma\}$ and $W(\lambda) = \#\{p_i > \lambda\}$.

(d) For any rejection region of interest $[0, \gamma]$, estimate pFDR as

$$p\widehat{FDR}(\gamma) = \frac{\hat{\pi}_0 \gamma}{\hat{F}_P(\gamma)\{1 - (1-\gamma)^m\}}.$$

(e) Estimate FDR as

$$\widehat{FDR}_\gamma = \frac{\hat{\pi}_0 \gamma}{\hat{F}_P(\gamma)}$$

Note: For details on choosing γ, see Section 9 of Storey (2002).

Box 3. Storey et al. (2004) procedure

(a) Estimate FDR using $\widehat{FDR}_\gamma$ from Box 2.

(b) Reject null hypotheses $p_i \leq t_\alpha(FDR_\gamma)$, $i = 1, \ldots, G$.

Box 4. Proposed Bayesian variable selection procedure # 1

(a) Set level to be α and fix a rejection region R.

(b) Fit model (2) - (5) using Markov Chain Monte Carlo (MCMC) methods.

(c) Based on the MCMC output, calculate $pp_i \equiv P(\gamma_i = 0|\hat{\beta}_i \in R)$, $i = 1, \ldots, G$.

(d) Let $pp_{(1)} \leq pp_{(2)} \leq \cdots \leq pp_{(G)}$ denote the sorted values of $pp_1, \ldots, pp_n$ in increasing order.

(e) Find $\hat{k} = \max\{1 \leq k \leq G : pp_{(k)} \leq \alpha k/G\}$; select variables $1, \ldots, G$.

Box 5. Proposed Bayesian variable selection procedure # 2

(a) Set level to be α and fix a rejection region R.

(b) Fit model (6) - (8) using Markov Chain Monte Carlo (MCMC) methods.

(c) Based on the MCMC output, calculate $pp_i \equiv P(\gamma_i = 0|\hat{\beta}_i \in R)$, $i = 1, \ldots, G$.

(d) Let $pp_{(1)} \leq pp_{(2)} \leq \cdots \leq pp_{(G)}$ denote the sorted values of $pp_1, \ldots, pp_n$ in increasing order.

(e) Find $\hat{k} = \max\{1 \leq k \leq G : pp_{(k)} \leq \alpha k/G\}$; select variables $1, \ldots, G$.

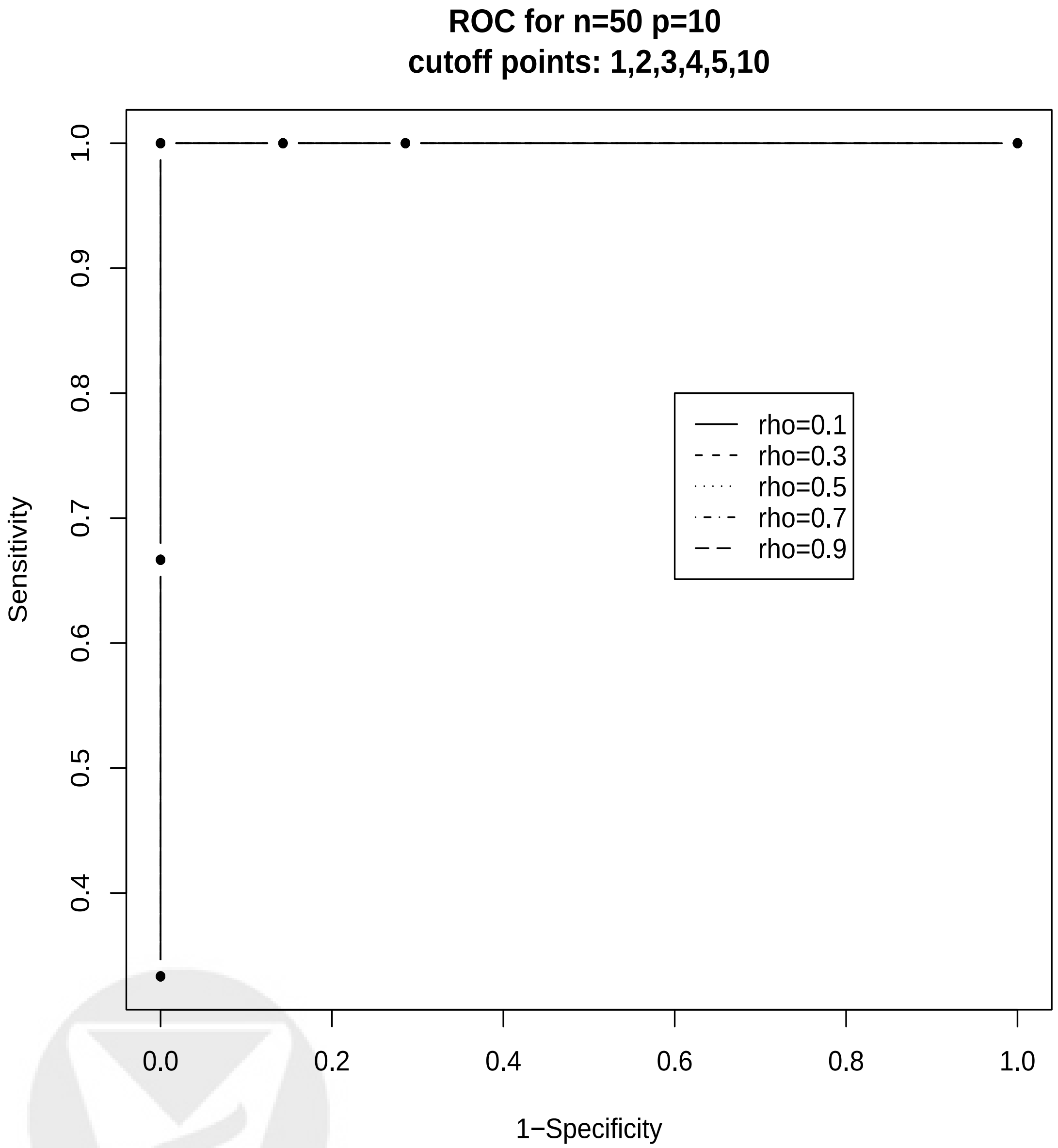

Figure 1: Plot of ROC curve for simulation setting when $n = 50$ and $p = 10$. Variables ranked univariately based on marginal posterior probability. ROC averaged across 250 simulations.

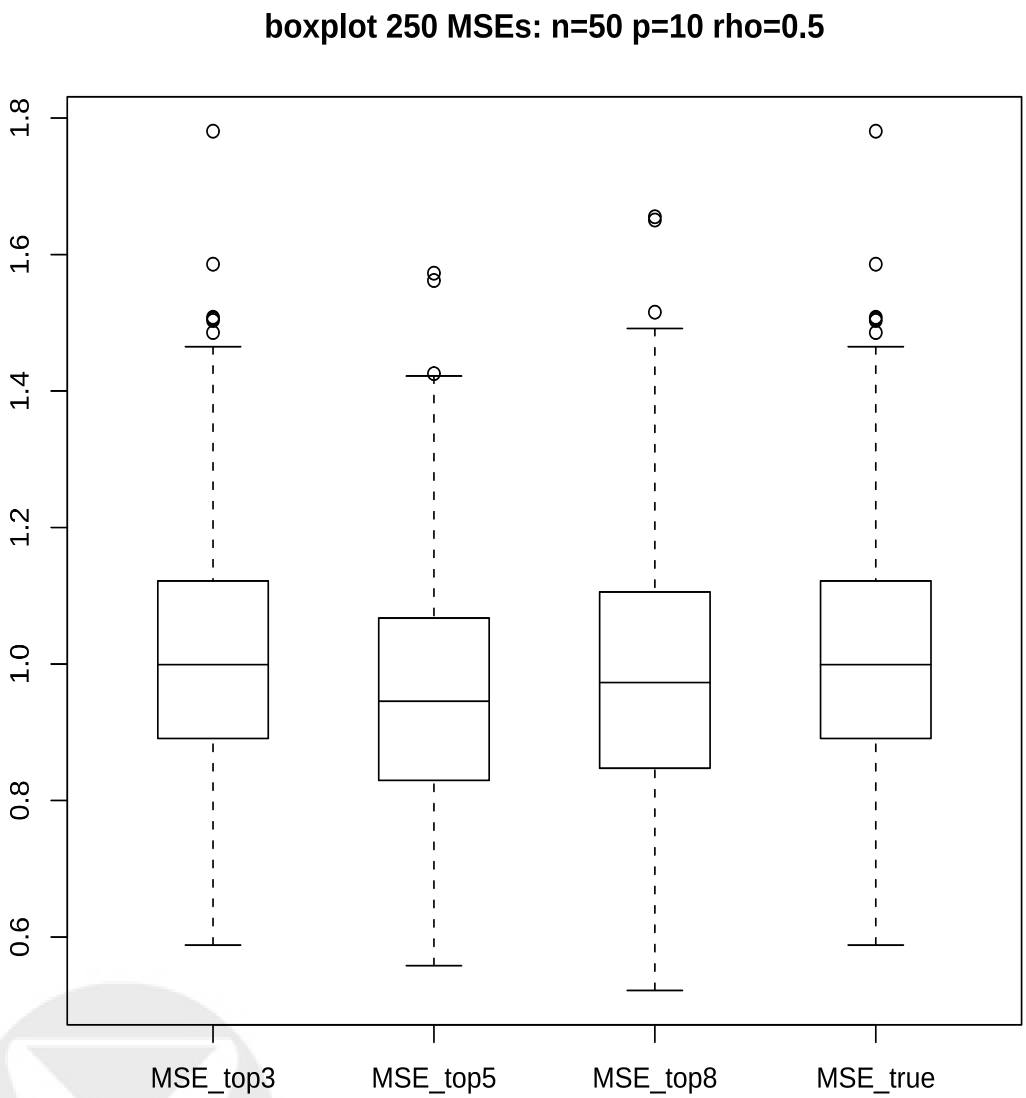

Figure 2: Mean squared errors based on taking top k variables ($k = 3, 5, 8$) and true MSE averaged across 250 simulations.

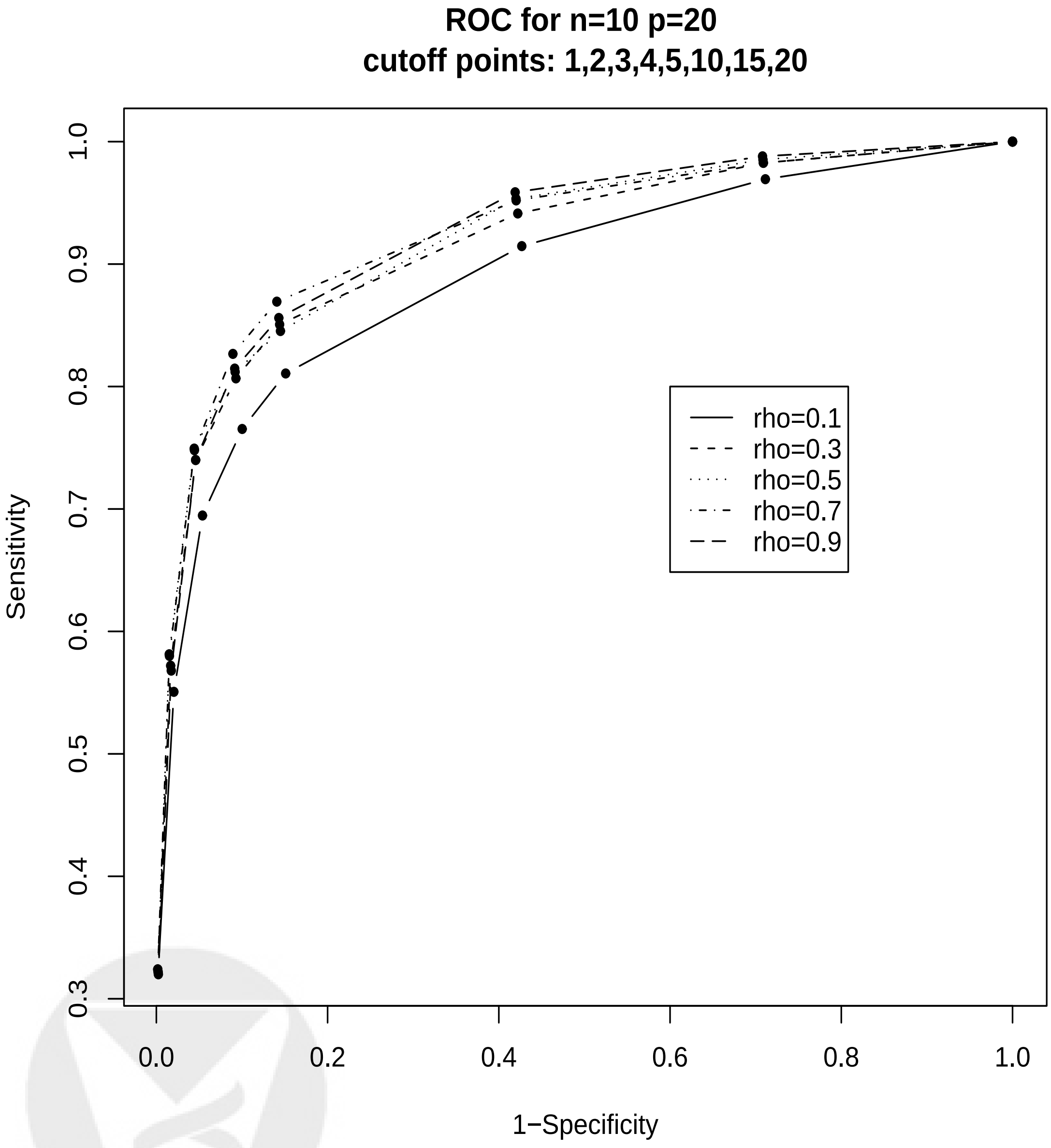

Figure 3: See caption to Figure 1.

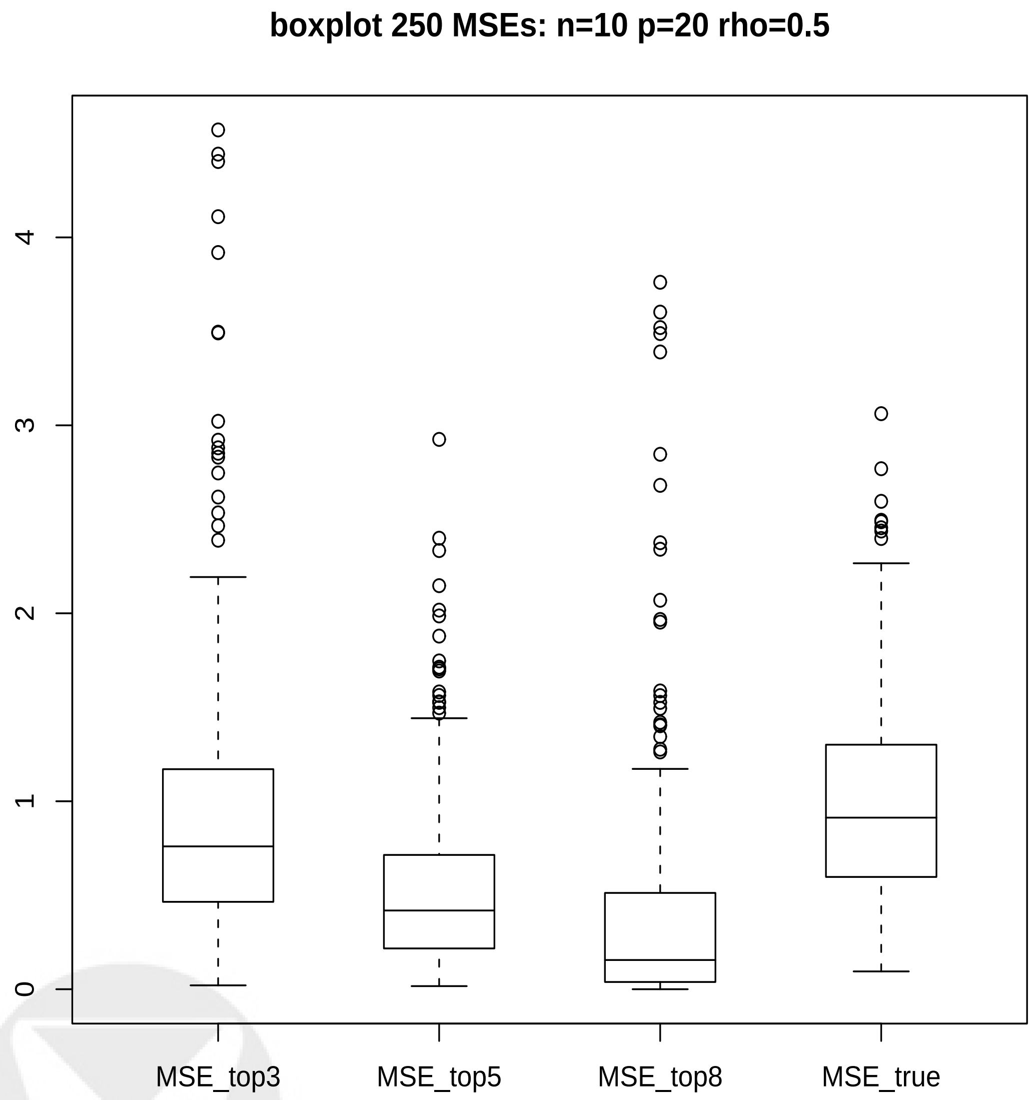

Figure 4: See averaged across 250 simulationscaption to Figure 2.