NONPARAMETRIC ESTIMATOR OF FALSE DISCOVERY RATE BASED ON BERNŠTEÍN POLYNOMIALS

Zhong Guan, Baolin Wu and Hongyu Zhao

Indiana University South Bend, University of Minnesota and Yale University School of Medicine

Abstract: Under a local dependence assumption about the p-values, an estimator of the proportion \( \pi_0 \) of true null hypotheses with a closed-form expression is derived based on the Bernšteín polynomial density estimation. A nonparametric estimator of false discovery rate (FDR) is then obtained. These estimators are proved to be consistent, and asymptotically unbiased and normal. Confidence intervals for \( \pi_0 \) and the FDR are also given. The usefulness of the proposed method is demonstrated through simulations and its application to a real microarray dataset.

Key words and phrases: Bernšteín polynomials, bioinformatics, density estimation, local dependence, false discovery rate, microarray, mixture model, multiple comparison.

1. Introduction

The statistical significance in multiple comparison problems has attracted attention of many authors. False discovery rate (FDR), first introduced by Benjamini and Hochberg (1995), is one measure of this statistical significance. Storey (2002a) and Storey and Tibshirani (2003) introduced the positive false discovery rate (pFDR) and proposed procedures for estimating FDR and pFDR with applications to DNA microarrays under the assumptions that the test statistics of the hypotheses are independent and dependent, respectively.

Let \( T \) be the test statistic for hypothesis \( H \). Denote the null and alternative hypotheses by \( H = 0 \) and \( H = 1 \), respectively. So \( \pi_0 \equiv \operatorname{Pr}(H = 0) \) is the proportion of true null hypotheses, and \( F_j(t) \equiv \operatorname{Pr}(T \leq t \mid H = j) \), \( j = 0, 1 \), are the null and alternative distribution of \( T \) respectively. Conduct this test \( n \) times simultaneously.
That is, consider testing hypotheses, $H_1, \ldots, H_n$, with test statistics $T_1, \ldots, T_n$. For each $i$, denote the null and alternative hypotheses by $H_i = 0$ and $H_i = 1$, respectively. Assume $Pr(H_i = 0) = \pi_0$ and $Pr(T_i \leq t | H_i = j) = F_j(t)$, $j = 0, 1$, for all $i$. The set of observed values, $t_1, \ldots, t_n$, of the test statistics $T_1, \ldots, T_n$ is treated as a sample from the mixture distribution of $T$

$$F(t) = \pi_0 F_0(t) + (1 - \pi_0) F_1(t). \tag{1.1}$$

Let $\Gamma$ be the common rejection region for all the tests. The notion of false non-discovery rate (FNR) was introduced by (Genovese and Wasserman, 2002). The following Bayesian interpretation of the pFDR and the positive false non-discovery rate (pFNR) can be found in (Storey, 2002a, 2003).

$$\text{pFDR} = Pr(H = 0 | T \in \Gamma) = \frac{\pi_0 Pr_{F_0}(T \in \Gamma)}{Pr_F(T \in \Gamma)},$$

$$\text{pFNR} = Pr(H = 1 | T \notin \Gamma) = 1 - \frac{\pi_0 Pr_{F_0}(T \notin \Gamma)}{Pr_F(T \notin \Gamma)}.$$  

The numerators $Pr_F(T \in \Gamma)$ and $Pr_F(T \notin \Gamma)$ of the above formulas can be estimated by empirical distribution. The probabilities $Pr_{F_0}(T \in \Gamma)$ and $Pr_{F_0}(T \notin \Gamma)$ can be obtained from the null distribution which usually is either known or can be estimated in some way such as using resampling methods. If $\pi_0$ can be estimated based on $t_1, \ldots, t_n$, then pFDR and pFNR are estimable. [Allison et al., (2002)] also used these quantities and modeled the distribution of the $p$-values from microarray data analysis by finite Beta mixture distribution. Note that the Type I error rate for each single test is $Pr(T \in \Gamma | H = 0)$ and the family-wise-error-rate (FWER) is $Pr(\bigcup_{i=1}^n (T_i \in \Gamma, H_i = 0))$ (see, for example, Westfall and Young, 1993; Ge et al., 2003).

The simplest situation in the application is that we know the parametric forms of both the null and alternative distributions $F_0$ and $F_1$. In this case, we can fit a parametric mixture model to these observed test statistics (Guan et al., 2004). Simulation studies showed that the model-based approach can significantly improve the pFDR and FDR estimation if the parametric model is correct.

In most applications, two other scenarios are more likely to occur. The first scenario is that the null distribution or at least its large sample approximation of the test statistics is known, but the alternative distribution is unknown. The second
scenario is the most difficult situation in which neither the null nor the alternative distributions are known. In this case, methods such as permutation procedures can be used to estimate the null distribution $F_0$ of the test statistics.

This paper focuses on the first scenario and assumes that both $F_0$ and $F_1$ are continuous and $F_0$ is known or can be estimated in some way such as permutation. We use $p$-values as the test statistics and, in this case, $F_0(t) = t, 0 \leq t \leq 1$. In the rest of the paper, $F_1$ is assumed to be continuous on $[0, 1]$ and $t_1, \ldots, t_n$ represent the $p$-values of the $n$ tests. In terms of densities, mixture model (1.1) with $F_0(t) = t$ can be written as

$$f(t) = \pi_0 + (1 - \pi_0)f_1(t).$$

In this case, if the common rejection region is \{ $p \leq p_0$ \}, then

$$\text{pFDR} = \frac{p_0 \pi_0}{F(p_0)} \quad \text{and} \quad \text{pFNR} = 1 - \frac{(1 - p_0)\pi_0}{1 - F(p_0)}.$$ 

Therefore, the key to the estimation of these quantities is the estimation of $\pi_0$. Write $b = \min_{t \in [0,1]} f_1(t)$. Clearly, $0 \leq b < 1$. In order that (1.2), as a nonparametric model, is identifiable, one has to assume that $b = 0$. Because, otherwise, for any $a \in [0, b]$, $\pi_0^* = \pi_0 + a(1 - \pi_0)$ and $f_1^*(t) = \{ f_1(t) - a \}/(1 - a)$ satisfy the model

$$f(t) = \pi_0^* + (1 - \pi_0^*)f_1^*(t).$$

Furthermore, the density $f_1$ is assumed to be continuous on $[0, 1]$ and $b = f_1(1) = 0$ so that $f(1) = \pi_0$ (Wu et al., 2006). Therefore, if $\hat{f}$ is a density estimate, then $\hat{f}(1)$ is an estimate of $\pi_0$.

The most commonly used kernel density estimate is subject to boundary effects at 0 and 1. In order to minimize the boundary effect of kernel density estimation for distribution with bounded support, one has to make boundary correction (Jones, 1993). Bernstein polynomial density estimate seems convenient for estimating $f(1)$ and has a closed-form expression. Let $\hat{F}$ denote the empirical distribution of $t_1, \ldots, t_n$. Storey and Tibshirani (2003) proposed using $\hat{\pi}_0 = \hat{g}(1)$ to estimate $\pi_0$ with $\hat{g}$ being the fitted spline to the data \{ $\hat{\pi}_0(\lambda) : \lambda = 0.01, 0.02, \ldots, 0.95$ \} where $\hat{\pi}_0(\lambda) = [1 - \hat{F}(\lambda)]/(1 - \lambda)$. Based on nonparametric maximum likelihood estimation of the density of $p$-value with restriction to convex decreasing densities, Langaas et al. (2005) proposed another smoothing method and showed that their method outperforms some existing estimators with respect to root-mean-squared error.

There are several approaches to the FDR estimation. Among many others,
Efron et al. (2001), Efron and Tibshirani (2002) and Efron (2003) proposed the empirical Bayes method which also uses the model (1.2), Guan et al. (2004) proposed a parametric method by assuming parametric forms of \( f_0 \) and \( f_1 \) in (1.2). Readers are referred to Wu et al. (2006) for an extensive comparison among these methods. Although not explored in Allison et al. (2002) directly, FDR could also be estimated by their method based on mixture model of Beta distributions.

Since any continuous function on \([0, 1]\) can be uniformly approximated with Bernstein polynomials (Bernstein 1912), Vitale (1975) proposed to use them to estimate unknown density function. Tenbusch (1994) extended this method to multidimensional situation. The rates of convergence of the posterior distribution for Bernstein polynomial prior were obtained by Ghosal (2001). The Bernstein polynomial and the \( k \)-th order Bernstein expansion of a function \( g(t) \) are defined as

\[
B_{j,k}(t) = \binom{k}{j} t^j (1-t)^{k-j}, \quad B_k g(t) = \sum_{j=0}^{k} g\left(\frac{j}{k}\right) B_{j,k}(t).
\]

One can estimate \( F \) and \( f \) by

\[
\hat{F}_k(t) = B_k \hat{F}(t) = \sum_{j=0}^{k} \hat{F}\left(\frac{j}{k}\right) B_{j,k}(t), \quad \hat{f}_k(t) = B_{k-1} \hat{f}(t) = \sum_{j=0}^{k-1} \hat{f}\left(\frac{j}{k-1}\right) B_{j,k-1}(t),
\]

respectively, where

\[
\hat{f}\left(\frac{j}{k-1}\right) = k \{ \hat{F}\left(\frac{j+1}{k}\right) - \hat{F}\left(\frac{j}{k}\right) \}, \quad \text{for } j = 0, \ldots, k - 1.
\]

In the estimation of \( f \), if \( k \) is chosen proportional to \( n^{2/5} \), then, for each fixed \( t \in (0, 1) \), the mean square error of \( \hat{f}_k(t) \) is proportional to \( n^{-4/5} \) (Vitale 1975).

In this paper a nonparametric method based on Bernstein polynomial density estimation is proposed. Simulation study and application to a microarray dataset are given in Section 4. The proofs of the main results are given in Appendix.

2. Estimators of \( \pi_0 \) and FDR and Asymptotic Results

Albeit a Bayesian interpretations of pFDR and pFNR have been used, it is convenient to work directly on the test statistics of the hypotheses. With a properly chosen \( 1 \leq r < k \), one can estimate \( \pi_0 \) by

\[
\tilde{\pi}_0 = \frac{1}{r} \sum_{l=1}^{r} \hat{f}_k(1 - \frac{l}{k}).
\]

\[\text{(2.1)}\]
If \( r = 1 \) then \( \tilde{\pi}_0 = \hat{f}_k(1 - 1/k) \approx f(1) = \pi_0 \) for large \( k \). As an average, this estimator may have smaller variance with larger \( r > 1 \). If \( r \) is too large, then the bias increases. Of course, the magnitude of the bias depends on \( k \) and \( f \) as well. Later in this paper, we shall develop a method to choose \( r \) and \( k \) to balance this bias-variance trade-off by minimizing a partial mean square error of \( \tilde{\pi}_0 \).

The following assumptions are needed for the asymptotic results about \( \tilde{\pi}_0 \):

**Assumption 1.** The test statistics \( T_1, \ldots, T_n \) satisfy a local dependence (LD1) of Chen and Shao (2005): for each \( T_i \), except \( n_i \) statistics \( T_{i_1}, \ldots, T_{i_{n_i}} \) all the other \( T_j \)'s are independent of \( T_i \). There exists an \( m \) independent of \( n \) so that \( \bar{n} = n^{-1} \sum_{i=1}^n n_i \leq m \). This is a generalization of the \( m \)-dependence.

**Assumption 2.** The second order partial derivative \( f_{uv}(s,t) = \partial^2 F_{uv}(s,t)/\partial s \partial t \) of the joint distribution function \( F_{uv}(s,t) \) of each pair \((T_u,T_v)\) is uniformly bounded by a constant independent of \((u,v)\).

Assumption 1 is usually satisfied for gene expression data since in the whole genome, each gene is likely to have interactions only with a limited number of other genes. A Glivenko-Cantelli lemma of (Yu, 1993) for dependent sequence ensures that Assumption 1 satisfies the weak dependence assumption made by Storey et al. (2004). More discussion on the dependence issue in the estimation of FDR can also be found in Langaas et al. (2005). Efron (2006) discusses the effect of correlation on the null distribution and FDR.

**Theorem 2.1.** Suppose that \( f(t) \) is continuously differentiable on \((0,1]\) with bounded derivative and \( f(1) = \pi_0 \). Then for each fixed \( r \), \( \tilde{\pi}_0 \) is an asymptotically unbiased estimator of \( \pi_0 \). Moreover, as \( k, n \to \infty \),

\[
|E(\tilde{\pi}_0) - \pi_0| = O(k^{-1}). \tag{2.2}
\]

If Assumptions 1 and 2 are fulfilled, then for each fixed \( r \),

\[
\lim_{k,n \to \infty} \frac{n \text{Var}(\tilde{\pi}_0)}{kh_k(r)} = \pi_0, \tag{2.3}
\]

where, for each \( r \geq 1 \),

\[
h_k(r) \equiv \sum_{j=0}^{k-1} \left\{ \frac{1}{r} \sum_{i=1}^r B_{j,k-1}(1-i/k) \right\}^2, \tag{2.4}
\]

\[
h(r) \equiv \lim_{k \to \infty} h_k(r) = \sum_{j=0}^{\infty} \left( \frac{1}{r} \sum_{l=1}^r \frac{BJ_{l}}{l^r} e^{-l} \right)^2 = O(r^{-3/2}). \tag{2.5}
\]
Furthermore, if \( k \) is of order \( n^{1/3} \), then
\[
E(\tilde{\pi}_0 - \pi_0)^2 = O(n^{-2/3}).
\] (2.6)

**Remark 2.1.** It should be noted that if the assumption \( f_1(1) = 0 \) is violated then \( \tilde{\pi}_0 \) is approximately conservative, i.e. \( E(\tilde{\pi}_0) > \pi_0 \) for large \( k \) and \( n \). Storey and Tibshirani (2003)’s estimate \( \hat{\pi}_0(\lambda) \) has the same property.

**Remark 2.2.** It is easy to see that
\[
h(r) = \frac{1}{r^2} \sum_{l=1}^{r} I_0(2l) e^{-2l} + \frac{2}{r^2} \sum_{1 \leq i < j \leq r} I_0(2\sqrt{ij}) e^{-i-j},
\] (2.7)
where \( I_0(x) \) is the modified first kind Bessel function \( I_0(\nu) \) with \( \nu = 0 \)
\[
I_0(x) = \sum_{j=0}^{\infty} \frac{(\frac{x}{2})^{2j}}{(j!)^2}.
\]
Clearly, \( h(r) \leq h(1) = 0.3085083 \).

**Remark 2.3.** The assumption that \( f'(1) \) is bounded can be violated in some cases. For example, let the test statistic \( T \) be \( N(0, 1) \) under \( H_0 \) and \( N(\mu, 1) \) under \( H_A \), with \( \mu > 0 \). The distribution function of the \( p \)-value of the one-sided test is \( F(t) = \pi_0 t + (1 - \pi_0) \{ 1 - \Phi[\Phi^{-1}(1 - t) - \mu] \} \), where \( \Phi \) is the distribution function of \( N(0, 1) \). Denote \( \varphi(t) = \Phi'(t) \). The density function and its derivative are respectively
\[
f(t) = F'(t) = \pi_0 + \sqrt{2\pi}(1 - \pi_0) \varphi(\mu) e^{\mu \Phi^{-1}(1-t)},
\]
\[
f'(t) = -2\pi \mu (1 - \pi_0) \varphi(\mu) e^{\Phi^{-1}(1-t) + \Phi^{-1}(1-t)}.
\]
Thus
\[
\lim_{t \to 1^-} f(t) = \pi_0, \quad \lim_{t \to 0^+} f'(t) = \lim_{t \to 1^-} f'(t) = -\infty.
\]
So \( f' \) is not bounded.

In the proof of Theorem 2.1, it is shown that \( \tilde{\pi}_0 \) is a sum of locally dependent random variables, the following asymptotic normality of \( \tilde{\pi}_0 \) is a consequence of the recent result of Chen and Shao (2005).

**Theorem 2.2.** Suppose that Assumptions 1 and 2 are fulfilled. If \( f(t) \) is continuously differentiable on \( (0, 1] \) with bounded derivative and \( f(1) = \pi_0 \), then for fixed \( r \) the estimator \( \tilde{\pi}_0 \) is asymptotically normal, i.e., as \( k, n \to \infty \) and \( k/n \to 0 \),
\[
\frac{\sqrt{n}(\tilde{\pi}_0 - E(\tilde{\pi}_0))}{\sqrt{kh_k(r)}} \xrightarrow{d} N(0, \pi_0).
\] (2.8)
Given a cutoff $p_0$ for the $p$-values, FDR can be estimated by

$$\hat{pFDR}(p_0) = \frac{p_0 \tilde{\pi}_0}{\hat{F}(p_0)}.$$  \hfill (2.9)

From the Glivenko-Cantelli lemma (Yu, 1993), Theorems 2.1 and 2.2, it follows that $\hat{pFDR}(p_0)$ is also consistent and asymptotically normal. One can construct a confidence interval for $\pi_0$ as follows. For a given confidence level $1 - \alpha$, let $z_{\alpha/2}$ be the upper $\alpha/2$ quantile of the standard normal distribution, i.e. $\Phi(z_{\alpha/2}) = 1 - \alpha/2$,

$$\Pr\left\{ \frac{\sqrt{n}}{k \tilde{\pi}_0 h_k(r)} | \tilde{\pi}_0 - \pi_0 | < z_{\alpha/2} \right\} \approx 1 - \alpha.$$

Therefore

$$\Pr\{ \tilde{\pi}_0_L(\alpha) < \pi_0 < \tilde{\pi}_0_U(\alpha) \} \approx 1 - \alpha,$$

where

$$\tilde{\pi}_0_L(\alpha) = \tilde{\pi}_0 - z_{\alpha/2} \sqrt{\frac{k h_k(r)}{n}} \tilde{\pi}_0 \quad \text{and} \quad \tilde{\pi}_0_U(\alpha) = \tilde{\pi}_0 + z_{\alpha/2} \sqrt{\frac{k h_k(r)}{n}} \tilde{\pi}_0.$$  \hfill (2.10)

Based on the confidence interval for $\pi_0$ one can obtain the confidence interval $(\hat{pFDR}_L, \hat{pFDR}_U)$ for pFDR with

$$\hat{pFDR}_L = \frac{p_0 \tilde{\pi}_0_L(\alpha)}{\hat{F}(p_0)}, \quad \hat{pFDR}_U = \frac{p_0 \tilde{\pi}_0_U(\alpha)}{\hat{F}(p_0)}.$$

One can also replace $\hat{F}(p_0)$ with $\hat{F}_k(p_0)$.

3. Choosing optimal $r$ and $k$

When the sample size $n$ is large, as in microarray data analysis, Assumptions 1 and 2 assure that the contribution made by covariances into the variance of the estimator $\tilde{\pi}_0$ is bounded above by a quantity independent of $(r, k)$ [see (A.17)]. In the proof of Theorem 2.1 (A.8) gives an estimate of the bias of $\tilde{\pi}_0$

$$B(r, k) = |E(\tilde{\pi}_0) - \pi_0| \leq \sum_{i=0}^3 R_{1i}(k, r).$$  \hfill (3.1)

So one can choose $r$ and $k$ by minimizing the partial mean square error

$$\text{pMSE}(r, k) = \left\{ \sum_{i=0}^3 R_{1i}(k, r) \right\}^2 + D(r, k),$$  \hfill (3.2)
where \( D(r, k) = (k\pi_0/n)h_k(r) \). One can estimate \( \text{pMSE}(r, k) \) by
\[
\hat{\text{pMSE}}(r, k) = \left\{ \sum_{i=0}^{3} \hat{R}_i(k, r) \right\}^2 + \hat{D}(r, k),
\]

(3.3)

where \( \hat{D}(r, k) = (k\hat{\pi}_0/n)h_k(r) \),

\[
\hat{R}_1(k, r) \approx (\frac{1}{2})^{1-i} \sum_{j=0}^{k-1} \hat{b}(j, k, r) |\hat{f}_k^{(j)}(\frac{j}{k-1})| |(\frac{j}{k-1})|^i, \quad i = 0, 1,
\]

(3.4)

\[
\hat{R}_2(k, r) \approx \sum_{j=0}^{k-1} \hat{b}(j, k, r) |\hat{f}_k^{(j)}(\frac{j}{k-1})| |1 - \frac{1}{k} - \frac{j}{k-1}|,
\]

(3.5)

\[
\hat{R}_3(k, r) \approx \frac{1}{k} |\hat{f}_k^{(1)}(1 - \frac{1}{k})|,
\]

(3.6)

\[
\hat{f}_k^{(j)}(t) = \mathbb{H}_{k-2}\hat{f}(t) = \sum_{j=0}^{k-2} \hat{f}_k^{(j)}(\frac{j}{k-2}) B_{j,k-2}(t),
\]

(3.7)

\[
\hat{f}_k^{(j)}(\frac{j}{k-2}) = (k-1)\{ \hat{f}_k^{(j+1)}(\frac{j+1}{k-1}) - \hat{f}_k^{(j)}(\frac{j}{k-1}) \}, \quad \text{for } j = 0, \ldots, k-2
\]

(3.8)

and \( \hat{b}(j, k, r) \) is defined by (A.2). The optimal \( \hat{r} \) and \( \hat{k} \) satisfy

\[
\hat{\text{pMSE}}(\hat{r}, \hat{k}) = \min\{\text{pMSE}(s, t), 1 \leq s < t < n\}.
\]

Intuitively, the larger the number \( k \) of bins, the larger the variance of \( \hat{\pi}_0 \). On the other hand, increasing the number \( r \) in (2.1) can reduce the variance of \( \hat{\pi}_0 \). Based on (2.4) and (2.5), \((k\pi_0/n)h_k(r)\) is an applicable measure of the dependence of variance of \( \hat{\pi}_0 \) upon \( r \) and \( k \). The upper bound estimate (3.1) is obtained by applying only the triangle inequality. So (3.2) is suitable for finding optimal \( r \) and \( k \). The R package, \textit{nFDR}, which implements the method of this paper is available on CRAN (the Comprehensive R Archive Network).

4. Simulation Studies and Application to Microarray Data

Comparison Study: In this simulation, \( n = 1000, \pi_0 = 0.25, 0.50, 0.75, \text{ and } 0.95 \) and \( B = 500 \) sets of \( p \)-values \( p_1, \ldots, p_n \) were simulated with \( p_i = \text{uniform}(0, 1) \) or \( \text{Beta}(1, 6) \). The proportion \( \pi_0 \) of true null hypotheses is estimated in four different ways. The results are summarized in Figure 1. The estimates are (1) \( \hat{\pi}_0 \) based on \((r^*, k^*)\), where \((r^*, k^*)\) is the minimizer of \( \text{pMSE}(r, k) \); (2) \( \hat{\pi}_0 \) based on \((\hat{r}, \hat{k})\); (3) \( \hat{\pi}_0^c \) which is estimated by function \textit{convest()} of R package \textit{limma} which implements the convex decreasing density method of Langaas et al. (2005); and (4) \( \hat{\pi}_0^p \) which is
estimated by R package qvalue using default setting which implements the smooth method described in Storey and Tibshirani (2003) (see also Storey 2003; Storey et al. 2004). This simulation shows that the \( \hat{\pi}_0 \)'s based on \( \hat{r}, \hat{k} \) have a variation close to but a little smaller bias than the ones based on \( r^*, k^* \). Thus the selected \( \hat{r} \) and \( \hat{k} \) perform well. The smooth method of Storey and Tibshirani (2003) has a larger variation and a larger bias than the proposed method. Except the differences in the biases, the proposed method has variation similar to the convex decreasing density method.

**Impact of Dependence:** In this simulation, \( n = 3000, m = 10 \) and \( B = 500 \). First, two-sample gene expression data \( \{x_{ij}, y_{ij} : i = 1, 2, \ldots, n; j = 1, 2, \ldots, N\} \), \( N = 10 \), are generated in three different ways. For each gene \( i \), the null hypothesis is \( H_i : \mu_x = \mu_y \). Then \( B \) sets of \( p \)-values \( p_i \) of two-sample \( t \)-test with the same variances are calculated based on \( x_{i1}, \ldots, x_{iN} \) and \( y_{i1}, \ldots, y_{iN} \).

1. **Independence:**
   \[
   x_{ij} = \mu_{ij}, \quad y_{ij} = \mu_{ij}^\prime + 3I\{i \leq (1 - \pi_0)n\}
   \]
   with \( \mu_{ij} \)'s and \( \mu_{ij}^\prime \)'s being i.i.d. \( N(0,1) \);

2. **Dependence Case 1:**
   \[
   x_{ij} = (-1)^i \mu_{v+j} + \varepsilon_{ij}, \quad y_{ij} = (-1)^i \mu_{v+j}^\prime + \varepsilon_{ij}^\prime + 2I\{i \leq (1 - \pi_0)n\},
   \]
   \( j = 1, \ldots, N, \ i = (v - 1)m + 1, \ldots, (v - 1)m + m, \ v = 1, \ldots, n/m \),
   where \( \varepsilon_{ij} \)'s and \( \varepsilon_{ij}^\prime \)'s are i.i.d. \( N(0,0.04^2) \) and, for each \( v \), \( \mu_{v+j} \)'s and \( \mu_{v+j}^\prime \)'s are i.i.d. \( N(0,1) \). In this case, the correlation between \( p \)-values for each pairs of genes in a group of \( m \) is about ±0.9983 by simulation. This dependence is similar to but has larger correlation than the dependence simulation of Storey et al. (2004).

3. **Dependence Case 2:**
   \[
   x_{ij} = \mu_{ij} + \varepsilon_j, \quad y_{ij} = \mu_{ij}^\prime + \varepsilon_j^\prime + 2I\{i \leq (1 - \pi_0)n\},
   \]
   \( j = 1, \ldots, N, \ i = 1, \ldots, n \),
   where \( \varepsilon_j \)'s and \( \varepsilon_j^\prime \)'s are i.i.d. \( N(0,0.25^2) \). In this case, the correlation between \( p \)-values for each pairs of genes is about 0.0581 by simulation. This is the so-called “general dependence” of Storey (2002b).
Figure 1: Simulation results for Beta distributed $p$-values. The true values of $\pi_0$ are 0.25, 0.50, 0.75, and 0.95.
The simulation $\tilde{\pi}_0$’s are summarized in Figure 2. The biases and standard deviations of the simulated $\tilde{\pi}_0$’s and the estimated coverage probabilities of the 95% confidence intervals for $\pi_0$ are given in Table 1. The results for $\pi_0 = 0.05$ are not shown in Figure 2. The above simulation studies show that the performance of the method is satisfactory for most cases. When dependence is present and $\pi_0$ is close to 1, the variance of $\tilde{\pi}_0$ may be underestimated so that the coverage is less than the nominal one. In applications to microarray data analysis, this can be overcome by eliminating many obvious non-significant and irrelevant genes using data preprocessing and filtering.

Table 1: Biases and standard deviations of the simulated $\tilde{\pi}_0$’s presented in Figure 2 and simulated coverage probabilities of 95% CI’s.

<table>
<thead>
<tr>
<th>$\pi_0$</th>
<th>0.05</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence</td>
<td>0.00084</td>
<td>0.00095</td>
<td>0.00096</td>
<td>-0.00077</td>
<td>-0.00230</td>
</tr>
<tr>
<td>Dependence 1</td>
<td>-0.00057</td>
<td>0.00045</td>
<td>-0.00044</td>
<td>-0.00079</td>
<td>-0.00123</td>
</tr>
<tr>
<td>Dependence 2</td>
<td>-0.00075</td>
<td>-0.00115</td>
<td>-0.00114</td>
<td>-0.00068</td>
<td>-0.00109</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\pi_0$</th>
<th>0.05</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence</td>
<td>0.0083</td>
<td>0.0194</td>
<td>0.0252</td>
<td>0.0297</td>
<td>0.0303</td>
</tr>
<tr>
<td>Dependence 1</td>
<td>0.0075</td>
<td>0.0164</td>
<td>0.0230</td>
<td>0.0299</td>
<td>0.0347</td>
</tr>
<tr>
<td>Dependence 2</td>
<td>0.0071</td>
<td>0.0150</td>
<td>0.0182</td>
<td>0.0219</td>
<td>0.0237</td>
</tr>
</tbody>
</table>

Leukemia Data: In large-scale microarray data analysis, there are usually thousands or tens of thousands of genes involved. It is practical to assume that genes in the same pathway have similar expression profiles and affect the system function in a synergistic way. The number of genes in a pathway is usually relatively small compared to the total number of genes in the data. The researchers are usually interested in identifying differentially expressed genes using certain type of test method. For each gene, a value of the test statistic is calculated based on sample observations on the expression levels. So test statistics such as $p$-values generated
Figure 2: Simulation results for independent and dependent $p$-values. The true values of $\pi_0$ are 0.25, 0.50, 0.75 and 0.95.
in microarray data analysis seem to satisfy the assumptions of this paper.

The leukemia gene expression dataset was reported in Golub et al. (1999). In this study there were \(N_1 = 47\) patients with Acute Lymphoblastic Leukemia (ALL) and \(N_2 = 25\) patients with Acute Myeloid Leukemia (AML). The mRNA levels of 7129 genes were measured for these \(N = 72\) samples. The same procedures as in Wu et al. (2006) were used to preprocess genes and calculate two sample \(t\)-test statistics. Permutations were used to obtain \(p\)-values for \(n = 3571\) remaining genes after data preprocessing and filtering. The histogram of the \(p\)-values (not shown here) indicates that the mixture model (1.2) seems valid. So the assumptions of the paper are likely not violated. Based on the expression data, correlation tests for the \(n(n-1)/2\) pairs of genes using Bonferroni adjusted FWER of 0.05 give an estimate of \(\bar{n} = 6.529\). The method of this paper results in \((\hat{r}, \hat{k}) = (18, 107)\) and \(\hat{\pi}_0 = 0.449\), which is low because many obvious non-significant and irrelevant genes have been eliminated by data preprocessing and filtering.

Based on the simulation study and Theorem 2.2, if possible, data preprocessing and filtering are usually recommended to have a smaller \(\pi_0\) and thus smaller variance of \(\hat{\pi}_0\). The corresponding confidence interval is (0.399, 0.498). Figure 3 shows the pFDR estimations and the 95% confidence intervals. For example, if \(p_0 = 1.808 \times 10^{-3}\) is a cutoff of the \(p\)-values, then there are about 600 genes which have smaller \(p\)-values and are claimed to be differentially expressed. The corresponding pFDR is \(4.944 \times 10^{-3}\).

**Acknowledgments**

The authors would like to thank the two co-editors, an associate editor, and two anonymous referees for their comments and suggestions which have greatly improved the presentation of the paper. They also thank Professor Terrence Speed for his valuable comments. This research was supported in part by startup funds from Indiana University South Bend and the Division of Biostatistics, University of Minnesota, and NIH grant GM59507 and NSF grant DMS 0241160.

**Appendix**

**Proof of Theorem 2.1**  By \(\hat{F}(t) = n^{-1} \sum_{i=1}^{n} I(T_i \leq t)\), it is easy to see that

\[
\hat{\pi}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_{ki},
\]  

(A.1)
Figure 3: pFDR Estimation for Golub et al. (1999) Data

\[ Y_{ki} = k \sum_{j=0}^{k-1} \tilde{b}(j, k, r)I\{ \frac{i}{k} < T_i \leq \frac{i+1}{k} \}, \quad \tilde{b}(j, k, r) = \frac{1}{r} \sum_{l=1}^{r} B_{j,k-1}(1 - \frac{l}{r}). \quad (A.2) \]

In the proof, the following results are useful. If \( |g_j| \leq M \), then for \( i \geq 1 \),

\[ \sum_{j=0}^{k-1} \tilde{b}^i(j, k, r)g_j = O(1) \quad (A.3) \]

and for \( v \geq i \geq 0 \) and \( v \geq 1 \),

\[ \sum_{j=0}^{v} B_{j,v}(t)j^{[i]} = v^{[i]}t^{i} \quad (A.4) \]

where \( j^{[i]} = j(j-1) \cdots (j-i+1) \) if \( j \geq i; = 0 \) otherwise. Define \( j^{[0]} = 1 \) for \( j > 0 \). Clearly, for each fixed \( k \), \( Y_{k1}, \ldots, Y_{kn} \) are identically distributed random variables
The Cauchy-Schwarz inequality implies
\[ \mu_k \equiv E(Y_k) = k \sum_{j=0}^{k-1} b(j, k, r) \Delta_{kj}, \tag{A.5} \]
\[ \sigma_k^2 \equiv \text{Var}(Y_k) = k^2 \sum_{j=0}^{k-1} b^2(j, k, r) \Delta_{kj} - \{E(Y_k)\}^2, \tag{A.6} \]
where \( \Delta_{kj} = F[(j + 1)/k] - F(j/k). \) Since \( f(1) = \pi_0, \ E(\tilde{\pi}_0) = E(Y_k) \) and \( \sum_{j=0}^{k-1} b(j, k, r) = 1, \) the absolute bias
\[ B(r, k) \equiv |E(\tilde{\pi}_0) - \pi_0| = |E(Y_k) - f(1)| = \left| \sum_{j=0}^{k-1} b(j, k, r) [k \Delta_{kj} - f(1)] \right|. \]
Applications of Taylor expansion imply that there exist \( \xi_{0j} \in (j/k, (j + 1)/k), \]
\( \xi_{1j} \in (j/k, j/(k - 1)), \xi_{2j} \in (j/(k - 1), 1 - 1/k) \) and \( \xi_3 \in (1 - 1/k, 1) \) such that
\[ k \Delta_{kj} = f(1) + \frac{1}{2k} f'(\xi_{0j}) = f\left(\frac{j}{k-1}\right) + \frac{1}{2k} f'(\xi_{0j}) - f'(\xi_{1j}) \frac{j}{k(k-1)} \]
\[ = f(1) + \frac{1}{2k} f'(\xi_{0j}) - f'(\xi_{1j}) \frac{j}{k(k-1)} - f'(\xi_{2j}) (1 - \frac{1}{k} - \frac{j}{k-1}) - f'(\xi_3) \frac{j}{k}. \tag{A.7} \]
Therefore
\[ B(r, k) \leq \sum_{i=0}^{3} R_{1i}(k, r), \tag{A.8} \]
where \( R_{1i}(k, r), \) \( i = 0, 1, 2, 3, \) are defined below. Since \( tf'(t) \) is bounded,
\[ k R_{1i}(k, r) \equiv \left(\frac{1}{2}\right)^{1-i} \sum_{j=0}^{k-1} b(j, k, r) |f'(\xi_{ij})| \left(\frac{j}{k-1}\right)^i \]
\[ \approx \left(\frac{1}{2}\right)^{1-i} \sum_{j=0}^{k-1} b(j, k, r) |f'(\frac{j}{k-1})| \left(\frac{j}{k-1}\right)^i = O(1), \quad i = 0, 1. \tag{A.9} \]
The Cauchy-Schwarz inequality implies
\[ R_{12}^2(k, r) \equiv \left\{ \sum_{j=0}^{k-1} b(j, k, r) |f'(\xi_{2j})| \right\}^2 \]
\[ \leq \sum_{j=0}^{k-1} b(j, k, r) \left\{ f'(\xi_{2j}) \right\}^2 \sum_{j=0}^{k-1} b(j, k, r) \left(1 - \frac{1}{k} - \frac{j}{k-1}\right)^2. \tag{A.10} \]
It follows from (A.4) that
\[ \sum_{j=0}^{k-1} b(j, k, r)(1 - \frac{i}{k} - \frac{j}{k-1})^2 = \frac{1}{r} \sum_{l=1}^{r} B_{j, k-l}(1 - \frac{i}{k})(1 - \frac{i}{k} - \frac{j}{k-1})^2 \]
\[ = \frac{1}{3k^2} (r^2 + 2 - \frac{r^2-1}{k-1}). \]  
(A.11)

It follows from (A.3), (A.10) and (A.11) that
\[ R_{12}(k, r) = \mathcal{O}(k^{-1}), \]  
(A.12)

\[ R_{13}(k, r) \equiv \sum_{j=0}^{k-1} b(j, k, r)f\left(\frac{f'(\xi_3)}{|k|^2}\right) \equiv \frac{1}{k} |f'(\xi_3)| \approx \frac{1}{k} |f'(1 - \frac{1}{k})|. \]  
(A.13)

Combination of (A.9) through (A.13) proves (2.2). For \( u \neq v \),
\[ \text{Cov}(Y_{ku}, Y_{kv}) = k^2 \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} b(i, k, r)b(j, k, r)\left\{(\Delta_{kij}^{(uv)} - \Delta_{kij}\Delta_{kji}\right\}, \]  
(A.14)

where
\[ \Delta_{kij}^{(uv)} = \text{Pr}\left(\frac{i}{k} < T_u < \frac{i+1}{k}, \frac{j}{k} < T_v \leq \frac{j+1}{k}\right) \]
\[ = F_{uv}\left(\frac{i+1}{k}, \frac{i}{k}\right) - F_{uv}\left(\frac{i}{k}, \frac{i+1}{k}\right) - F_{uv}\left(\frac{j}{k}, \frac{j+1}{k}\right) + F_{uv}\left(\frac{j+1}{k}, \frac{j}{k}\right). \]  
(A.15)

By the mean value theorem, we have that for some \( \tau_{ki} \in (i/k, (i+1)/k) \), \( \zeta_i^{(uv)} \in (i/k, (i+1)/k) \), and \( \eta_j^{(uv)} \in (j/k, (j+1)/k) \)
\[ k\Delta_{ki} = f(\tau_{ki}), \quad k^2\Delta_{kij}^{(uv)} = f_{uv}(\zeta_i^{(uv)}, \eta_j^{(uv)}). \]  
(A.16)

Since \( f \) and \( f_{uv} \) are bounded, it follows from (A.14) and (A.16) that there exists constant \( C > 0 \) such that for \( u \neq v \)
\[ |\text{Cov}(Y_{ku}, Y_{kv})| \leq \max_{i,j} \left\{ f_{uv}(\zeta_i^{(uv)}, \eta_j^{(uv)}) + f(\tau_{ki})f(\tau_{kj}) \right\} \left[ \sum_{i=0}^{k-1} b(i, k, r) \right]^2 \leq C \]
and
\[ |\text{Var}(\bar{s}_0) - \frac{1}{n} \sigma_k^2| = \frac{1}{n^2} \sum_{u \neq v} \text{Cov}(Y_{ku}, Y_{kv}) \leq \frac{(n-1)C}{n} \leq \frac{m-1}{n} C. \]  
(A.17)

From (A.3) it follows that
\[ kR_{2i}(k, r) \equiv \sum_{j=0}^{k-1} b^2(j, k, r)|f'(\xi_{ij})|\left(\frac{i}{k-1}\right)^i = \mathcal{O}(1), \; i = 0, 1, \]  
(A.18)
Application of the Cauchy-Schwarz inequality again gives
\[
R_{22}^2(k, r) \equiv \left\{ \sum_{j=0}^{k-1} \bar{b}^2(j, k, r) \left| f'(\xi_2j)\right| \left| 1 - \frac{1}{k} - \frac{j}{k+1} \right| \right\}^2
\leq \sum_{j=0}^{k-1} \bar{b}^3(j, k, r) \left| f'(\xi_2j)\right|^2 \sum_{j=0}^{k-1} \bar{b}(j, k, r) \left( 1 - \frac{1}{k} - \frac{j}{k+1} \right)^2
= \frac{1}{3k^2} \left\{ \sum_{j=0}^{k-1} \bar{b}^3(j, k, r) \left| f'(\xi_2j)\right|^2 \right\} \left\{ r^2 + 2 - \frac{r^2-1}{k+1} \right\} = O(r^2k^{-2}), \quad (A.19)
\]
\[
R_{23}(k, r) \equiv \frac{1}{k} \left| f'(\xi_3)\right| h_k(r) = O(k^{-1}). \quad (A.20)
\]
From these, it follows that
\[
\left| k \sum_{j=0}^{k-1} \bar{b}^2(j, k, r) \Delta_{kj} - f(1)h_k(r) \right| = O(k^{-1}). \quad (A.21)
\]
This, \((2.5)\), \((A.6)\) and \((A.17)\) imply
\[
\frac{\text{Var}(Y_{ki})}{kh_k(r)} = \pi_0 - \frac{\mu_k}{kh_k(r)} + O(k^{-2}) = \pi_0 + O(k^{-1}),
\]
\[
\frac{n \text{Var}(\tilde{\pi}_0)}{kh_k(r)} = \frac{\text{Var}(Y_{ki})}{kh_k(r)} + O(k^{-1}).
\]
Consequently
\[
\lim_{k, n \to \infty} \frac{n \text{Var}(\tilde{\pi}_0)}{kh_k(r)} = \lim_{k \to \infty} \frac{\text{Var}(Y_{ki})}{kh_k(r)} = \pi_0, \quad (A.22)
\]
\[
\text{E}(\tilde{\pi}_0 - \pi_0)^2 = O(k^{-2}) + O\left(\frac{\xi}{n}\right).
\]
If \(k\) is of order \(n^{1/3}\), then \((2.6)\) follows.
Let \(X_1, \ldots, X_l, Y_1, \ldots, Y_l\) be iid Poisson r.v.’s with mean 1. Then
\[
I_0(2l)e^{-2l} = \sum_{j=0}^{l} \left( \frac{l^j}{j!} e^{-l} \right)^2 = \text{Pr}\{ \sum_{i=1}^{l} X_i - \sum_{i=1}^{l} Y_i = 0 \}. \quad (A.23)
\]
The local limit theorem \(\text{(see Pages 187-188 of \cite{Petrov1975})}\) ensures
\[
\lim_{l \to \infty} \sqrt{l} \text{Pr}\{ \sum_{i=1}^{l} (X_i - Y_i) = 0 \} = \frac{1}{2\sqrt{\pi}}.
\]
From this it follows that there are constants \(0 < C_1 < C_2\) such that
\[
\frac{C_1}{\sqrt{l}} \leq I_0(2l)e^{-2l} \leq \frac{C_2}{\sqrt{l}}, \quad \text{for } l \geq 1, \quad (A.24)
\]
\[ I_0(2\sqrt{ij})e^{-ij} = I_0(2\sqrt{ij})e^{-2\sqrt{ij}}e^{-\sqrt{ij}^2} \begin{cases} \geq C_1(ij)^{-1/4}e^{-(\sqrt{i}-\sqrt{j})^2}, \\ \leq C_2(ij)^{-1/4}e^{-(\sqrt{i}-\sqrt{j})^2}. \end{cases} \quad (A.25) \]

Combining (2.7), (A.23) – (A.25), one obtains
\[
h(r) \leq C_2 \left\{ \frac{1}{r^2} \sum_{l=1}^{r} l^{-1/2} + \frac{2}{r^2} \sum_{1 \leq i < j \leq r} e^{-(\sqrt{i}-\sqrt{j})^2} (ij)^{-1/4} \right\} 
\leq \frac{2C_2}{r^{3/2}} \left\{ \int_0^1 t^{-1/2} dt + \frac{4}{\sqrt{r}} \int_0^{\sqrt{r}} dv \int_0^{v} \sqrt{ue^{-(u-v)^2}} dv \right\} \leq C_2^r r^{-3/2}.
\]

Similarly, \( h(r) \geq C'_2 r^{-3/2} \). The proof of Theorem 2.1 is complete.

**Proof of Theorem 2.2** Define
\[
\xi_i = \frac{Y_{ki} - E(\tilde{\pi}_0)}{n\sqrt{\text{Var(\tilde{\pi}_0)}}}, \quad i = 1, 2, \ldots, n.
\]
So \( \xi_i \) has mean zero and \( W = \sum_{i=1}^{n} \xi_i \) has variance \( \text{Var}(W) = 1 \). By Assumption 1, \( \xi_1, \ldots, \xi_n \) are also LD1 random variables. For each \( i \), let \( \eta_i \) be the sum of all the random variables \( \xi_i \), \ldots, \( \xi_{i_n} \) which are not independent of \( \xi_i \). By Theorem 3.4 of Chen and Shao (2005), we have
\[
\sup_x |\Pr(W \leq x) - \Phi(x)| \leq 2\delta^{1/2} \quad \text{ (A.26)}
\]
where
\[
\delta = 4E\left| \sum_{i=1}^{n} \{\xi_i \eta_i - E(\xi_i \eta_i)\} \right| + \sum_{i=1}^{n} E(|\xi_i \eta_i^2|) \equiv \delta_1 + \delta_2,
\]
\[
\delta_1 = 4E\left| \sum_{i=1}^{n} (\xi_i \eta_i - E(\xi_i \eta_i)) \right| = 4E\left| \sum_{i=1}^{n} \sum_{j=1}^{n_i} (\xi_i \xi_{ij} - E(\xi_i \xi_{ij})) \right|
\leq \frac{4}{n^2 \text{Var(\tilde{\pi}_0)}} \left\{ E\left| \sum_{i=1}^{n} \sum_{j=1}^{n_i} [Y_{ki}Y_{kij} - E(Y_{ki}Y_{kij})] \right| 
+ E\left| \sum_{i=1}^{n} \sum_{j=1}^{n_i} E(Y_{ki})[Y_{kij} - E(Y_{kij})] \right| + E\left| \sum_{i=1}^{n} \sum_{j=1}^{n_i} E(Y_{kij})[Y_{ki} - E(Y_{ki})] \right| \right\}
\equiv \delta_{11} + \delta_{12} + \delta_{13}.
\]
It is easy to see that there exists $C_3 > 0$ such that
\[
E[|Y_{ki}Y_{kj} - E(Y_{ki}Y_{kj})|] \\
\leq k^2 \sum_{u=0}^{k-1} \sum_{v=0}^{k-1} \tilde{b}(u, k, r)\tilde{b}(v, k, r)E[I\{\frac{u}{k} < T_i \leq \frac{u+1}{k}; \frac{v}{k} < T_j \leq \frac{v+1}{k}\} - \Delta_{kuv}^{(ij)}] \\
= 2k^2 \sum_{u=0}^{k-1} \sum_{v=0}^{k-1} \tilde{b}(u, k, r)\tilde{b}(v, k, r)\Delta_{kuv}^{(ij)}(1 - \Delta_{kuv}^{(ij)}) \leq C_3.
\]
Therefore
\[
\delta_{11} = \frac{4}{n^2 \Var(\hat{\pi}_0)} \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \sum_{u=0}^{k-1} \sum_{v=0}^{k-1} \tilde{b}(u, k, r)\tilde{b}(v, k, r)E[I\{\frac{u}{k} < T_i \leq \frac{u+1}{k}; \frac{v}{k} < T_j \leq \frac{v+1}{k}\} - \Delta_{kuv}^{(ij)}] \\
\leq \frac{4C_3m}{n \Var(\hat{\pi}_0)} = \mathcal{O}\left(\frac{1}{khk(r)}\right) = \mathcal{O}\left(\frac{r^{3/2}}{k}\right). \quad (A.27)
\]
Similarly there exists $C_4 > 0$ such that for any $j$
\[
E[|Y_{kj} - E(Y_{kj})|] \leq k \sum_{u=0}^{k-1} \tilde{b}(u, k, r)E[I\{\frac{u}{k} < T_j \leq \frac{u+1}{k}\} - \Delta_{ku}] \\
= 2k \sum_{u=0}^{k-1} \tilde{b}(u, k, r)\Delta_{ku}(1 - \Delta_{ku}) \leq C_4.
\]
Thus
\[
\delta_{12} = \frac{4}{n^2 \Var(\hat{\pi}_0)} \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \sum_{u=0}^{k-1} \tilde{b}(u, k, r)\mu_k E[|Y_{ki} - E(Y_{ki})|] \\
\leq \frac{4C_4m\mu_k}{n \Var(\hat{\pi}_0)} = \mathcal{O}\left(\frac{1}{khk(r)}\right) = \mathcal{O}\left(\frac{r^{3/2}}{k}\right). \quad (A.28)
\]
Similarly $\delta_{13} = \mathcal{O}(r^{3/2}/k)$. So $\delta_1 = \mathcal{O}(r^{3/2}/k)$.
\[
\delta_2 = \frac{\sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \sum_{u=1}^{n_i \Var(\hat{\pi}_0)} E[|Y_{ki} - \mu_k||Y_{kj} - \mu_k|^2]}{n^3 \Var(\hat{\pi}_0)} = \mathcal{O}(\frac{r^{3/2}}{k/n}).
\]
Thus, by (A.26), for fixed $r$ as $k, n \to \infty$ and $k/n \to 0$, $W \xrightarrow{d} N(0,1)$. The asymptotic normality (2.8) follows from this and (2.3).
References


Department of Mathematical Sciences, Indiana University South Bend, South Bend, IN 46634, U.S.A.

E-mail: zguan@iusb.edu

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, U.S.A.

E-mail: baolin@biostat.umn.edu

Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, U.S.A.

E-mail: hongyu.zhao@yale.edu