Nonparametric Estimator of False Discovery Rate Based on Bernšteĭn Polynomials

Guan, Zhong; Wu, Baolin; Zhao, Hongyu

To cite this article: Guan, Zhong, et al. "Nonparametric Estimator of False Discovery Rate Based on Bernšteĭn Polynomials." *Statistica Sinica*, vol. 18, 2008, pp. 905–23.

Statistica Sinica **18**(2008), 905-923

NONPARAMETRIC ESTIMATOR OF FALSE DISCOVERY RATE BASED ON BERNŠTEĬN 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 π_0 of true null hypotheses, having a closed-form expression, is derived based on Bernšteĭn polynomial density estimation. A nonparametric estimator of false discovery rate (FDR) is then obtained. These estimators are proved to be consistent, asymptotically unbiased, and normal. Confidence intervals for π_0 and the FDR are also given. The usefulness of the proposed method is demonstrated through simulations and its application to a microarray dataset.

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

1. Introduction

Statistical significance in multiple comparison problems has attracted the attention of many authors. The 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 $\pi_0 \equiv \Pr(H = 0)$ is the proportion of true null hypotheses, and $F_j(t) \equiv \Pr(T \le t \,|\, H = j)$, $j = 0, 1$, are the null and alternative distribution of T, respectively. Consider testing n hypotheses, $H_1, \ldots, H_n$, with test statistics T_1, ..., 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 $\Pr(T_i \le t \,|\, H_i = j) = F_j(t)$, $j = 0, 1$, 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 Γ 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 terms $\Pr_F(T \in \Gamma)$ and $\Pr_F(T \notin \Gamma)$ above can be estimated from the data. 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 known or can be estimated in some way, such as by using resampling methods. If π_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 a 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) and Ge, Dudoit and Speed (2003)).

The simplest situation occurs when 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 the observed test statistics (Guan, Wu and Zhao (2004)). Simulation studies have shown that the model-based approach can significantly improve pFDR and FDR estimation if the parametric model is correct.

In most applications, two other scenarios are more likely to occur. The first has the null distribution, or at least its large sample approximation, of the test statistics as known, while the alternative distribution is unknown. The other is the more 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 of the test statistics.

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

$$f(t) = \pi_0 + (1 - \pi_0) f_1(t). \tag{1.2}$$

In this case, if the common rejection region is $\{p \le p_0\}$, 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 π_0. Write $b = \min_{t\in[0,1]} f_1(t)$. Clearly, $0 \le b < 1$. In order that (1.2), as a nonparametric model, is identifiable, one has to assume that $b = 0$; 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 $[0,1]$ with $b = f_1(1) = 0$, so that $f(1) = \pi_0$ Wu, Guan and Zhao (2006). Therefore, if $\hat{f}$ is a density estimate, $\hat{f}(1)$ is an estimate of π_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 a boundary correction (Jones (1993)). A Bernšteĭn 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 π_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-values, with restriction to convex decreasing densities, Langaas, Lindqvist and Ferkingstad (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 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 method that assumed 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 directly in Allison et al. (2002), FDR could also be estimated by their method based on a mixture model of Beta distributions.

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

$$B_{j,k}(t) = \binom{k}{j} t^j (1-t)^{k-j}, \quad \mathbb{B}_k g(t) = \sum_{j=0}^{k} g\Big(\frac{j}{k}\Big) B_{j,k}(t).$$

One can estimate F and f by

$$\hat{F}_k(t) = \mathbb{B}_k \hat{F}(t) = \sum_{j=0}^{k} \hat{F}\Big(\frac{j}{k}\Big) B_{j,k}(t), \quad \hat{f}_k(t) = \mathbb{B}_{k-1} \hat{f}(t) = \sum_{j=0}^{k-1} \hat{f}\Big(\frac{j}{k-1}\Big) B_{j,k-1}(t),$$

respectively, where

$$\hat{f}\Big(\frac{j}{k-1}\Big) = k\Big\{\hat{F}\Big(\frac{j+1}{k}\Big) - \hat{F}\Big(\frac{j}{k}\Big)\Big\}, \quad \text{for } j = 0, \ldots, k-1. \tag{1.3}$$

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)).

We propose a nonparametric method based on Bernšteĭn polynomial density estimation. Simulation study and an application to a microarray dataset are carried out in Section 4. The proofs of the main results are given in the Appendix.

2. Estimators of π_0 and FDR and Asymptotic Results

Albeit Bayesian interpretations of pFDR and pFNR have been used, it is convenient to work directly with the test statistics of the hypotheses. With a properly chosen $1 \le r < k$, one can estimate π_0 by

$$\tilde{\pi}_0 = \frac{1}{r}\sum_{l=1}^{r} \hat{f}_k\Big(1 - \frac{l}{k}\Big). \tag{2.1}$$

If $r = 1$ then $\tilde{\pi}_0 = \hat{f}_k\big(1 - 1/k\big) \approx f(1) = \pi_0$ for large k. On average, this estimator has smaller variance for larger $r > 1$, but for larger r, bias increases. Of course, the magnitude of the bias depends on k and f as well. Later in this paper, we develop a method to choose r and k to balance bias-variance trade-o 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 the local dependence (LD1) of Chen and Shao (2005): for each T_i, except for n_i statistics $T_{i_1}, \ldots, T_{i_{n_i}}$ all other T_j's are independent of T_i. There exists an m independent of n so that $\bar{n} \equiv n^{-1}\sum_{i=1}^{n} n_i \le m$. This is a generalization of m-dependence.

Assumption 2. The 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 sequences 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 1. *Suppose $f(t)$ is continuously di erentiable on $(0,1]$ with a bounded derivative, and that $f(1) = \pi_0$. Then for each fixed r, $\tilde{\pi}_0$ is an asymptotically unbiased estimator of π_0. Moreover, as $k, n \to \infty$,*

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

If Assumptions 1 and 2 hold, then for each fixed r,

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

where, for each $r \geq 1$,

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

$$h(r) \equiv \lim_{k\to\infty} h_k(r) = \sum_{j=0}^{\infty} \left(\frac{1}{r} \sum_{l=1}^{r} \frac{l^j}{j!} e^{-l} \right)^2 = \mathcal{O}(r^{-\frac{3}{2}}). \tag{2.5}$$

Furthermore, if k is of order $n^{1/3}$, then

$$\mathrm{E}(\tilde{\pi}_0 - \pi_0)^2 = \mathcal{O}(n^{-\frac{2}{3}}). \tag{2.6}$$

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. The Storey and Tibshirani (2003) estimate $\hat{\pi}_0(\lambda)$ has the same property.

Remark 2.2. It is easy to see that

$$h(r) = \frac{1}{r^2} \sum_{l=1}^{r} I_0(2l) e^{-2l} + \frac{2}{r^2} \sum_{1\leq i<j\leq r} I_0(2\sqrt{ij}) e^{-i-j}, \tag{2.7}$$

where $I_0(x)$ is the modified first kind Bessel function $I_\nu(x)$ with $\nu = 0$:

$$I_0(x) = \sum_{j=0}^{\infty} \frac{\left(\frac{x}{2}\right)^{2j}}{(j!)^2}.$$

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 Φ is the distribution

function of $N(0,1)$. Write $\varphi(t) = \Phi'(t)$. The density function and its derivative are, respectively,

$$f(t) = F'(t) = \pi_0 + \sqrt{2\pi}(1-\pi_0)\varphi(\mu)e^{\mu\Phi^{-1}(1-t)},$$

$$f'(t) = -2\pi\mu(1-\pi_0)\varphi(\mu)e^{\Phi^{-1}(1-t)[\mu+\Phi^{-1}(1-t)]}.$$

Thus $\lim_{t\to 1-} f(t) = \pi_0$ and $\lim_{t\to 0+} f'(t) = \lim_{t\to 1-} f'(t) = -\infty$, 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 *hold. If $f(t)$ is continuously di erentiable on $(0,1]$ with bounded derivative and $f(1) = \pi_0$, then for fixed r, as $k, n \to \infty$ and $k/n \to 0$,*

$$\frac{\sqrt{n}\{\tilde{\pi}_0 - \mathrm{E}\,(\tilde{\pi}_0)\}}{\sqrt{kh_k(r)}} \xrightarrow{d} N(0, \pi_0). \tag{2.8}$$

Given a cuto p_0 for the p-values, FDR can be estimated by

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

From the Glivenko-Cantelli lemma of Yu (1993), Theorems 2.1 and 2.2, it follows that $\widehat{\mathrm{pFDR}}(p_0)$ is also consistent and asymptotically normal. One can construct a confidence interval for π_0 as follows. For a given confidence level $1-$, let $z_{\ /2}$ be the upper $/2$ quantile of the standard normal distribution, so

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

Therefore $\Pr\{\tilde{\pi}_{0_L}(\ ) < \pi_0 < \tilde{\pi}_{0_U}(\ )\} \approx 1-$, where

$$\tilde{\pi}_{0_L}(\ ) = \tilde{\pi}_0 - z_{\ /2}\sqrt{\frac{k}{n}h_k(r)\tilde{\pi}_0} \quad \text{and} \quad \tilde{\pi}_{0_U}(\ ) = \tilde{\pi}_0 + z_{\ /2}\sqrt{\frac{k}{n}h_k(r)\tilde{\pi}_0}\ . \tag{2.10}$$

Based on the confidence interval for π_0 one can obtain the confidence interval $(\widehat{\mathrm{pFDR}}_L, \widehat{\mathrm{pFDR}}_U)$ for pFDR with

$$\widehat{\mathrm{pFDR}}_L = \frac{p_0\tilde{\pi}_{0_L}(\ )}{\hat{F}(p_0)}, \quad \widehat{\mathrm{pFDR}}_U = \frac{p_0\tilde{\pi}_{0_U}(\ )}{\hat{F}(p_0)}.$$

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

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 to 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 $\tilde{\pi}_0$:

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

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

$$\mathrm{pMSE}(r,k) = \left\{ \sum_{i=0}^{3} R_{1i}(k,r) \right\}^2 + D(r,k), \tag{3.2}$$

where $D(r,k) = (k\pi_0/n)h_k(r)$. One can estimate $\mathrm{pMSE}(r,k)$ by

$$\widehat{\mathrm{pMSE}}(r,k) = \left\{ \sum_{i=0}^{3} \hat{R}_{1i}(k,r) \right\}^2 + \hat{D}(r,k), \tag{3.3}$$

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

$$k\hat{R}_{1i}(k,r) \approx \Big(\frac{1}{2}\Big)^{1-i} \sum_{j=0}^{k-1} \bar{b}(j,k,r)\ \hat{f}_k'\Big(\frac{j}{k-1}\Big)\ \Big(\frac{j}{k-1}\Big)^i,\ i = 0,1, \tag{3.4}$$

$$\hat{R}_{12}(k,r) \approx \sum_{j=0}^{k-1} \bar{b}(j,k,r)\ \hat{f}_k'\Big(\frac{j}{k-1}\Big)\ \ 1 - \frac{1}{k} - \frac{j}{k-1}\ , \tag{3.5}$$

$$\hat{R}_{13}(k,r) \approx \frac{1}{k}\ \hat{f}_k'\Big(1 - \frac{1}{k}\Big)\ , \tag{3.6}$$

$$\hat{f}_k'(t) = \mathbb{B}_{k-2}\hat{f}(t) = \sum_{j=0}^{k-2} \hat{f}'\Big(\frac{j}{k-2}\Big) B_{j,k-2}(t), \tag{3.7}$$

$$\hat{f}'\Big(\frac{j}{k-2}\Big) = (k-1)\Big\{ \hat{f}_k\Big(\frac{j+1}{k-1}\Big) - \hat{f}_k\Big(\frac{j}{k-1}\Big) \Big\}, \quad \text{for } j = 0, \ldots, k-2, \tag{3.8}$$

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

$$\widehat{\mathrm{pMSE}}(\hat{r},\hat{k}) = \min\{\widehat{\mathrm{pMSE}}(s,t),\ 1 \le s < t < n\}.$$

Intuitively, the larger the number k of bins, the larger the variance of $\tilde{\pi}_0$. On the other hand, increasing the number r in (2.1) can reduce the variance of $\tilde{\pi}_0$. Based

on (2.4) and (2.5), $(k\pi_0/n)h_k(r)$ is an applicable measure of the dependence of variance of $\tilde{\pi}_0$ upon r and k. The upper bound estimate (3.1) is obtained by applying the triangle inequality. Thus (3.2) is suitable for finding optimal r and k. The R package, `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, we took n =1,000, $\pi_0 = 0.25, 0.50, 0.75,$ and 0.95, and $B = 500$ sets of p-values $p_1, \ldots, p_n$ were simulated with p_i uniform(0, 1) or Beta$(1, 6)$. The proportion π_0 of true null hypotheses is estimated in four different ways: (1) $\tilde{\pi}_0^*$ is based on (r^*, k^*), where (r^*, k^*) is the minimizer of pMSE(r, k); (2) $\tilde{\pi}_0$ is based on $(\hat{r}, \hat{k})$; (3) $\hat{\pi}_0^c$ is estimated by function `convest()` of R package `limma`, which implements the convex decreasing density method of Langaas et al. (2005); and (4) $\hat{\pi}_0^q$ is estimated by the R package `qvalue` using a default setting that implements the smooth method described in Storey and Tibshirani (2003) (see also Storey (2003), Storey, Taylor and Siegmund (2004)). Simulation shows that the $\tilde{\pi}_0$'s based on $(\hat{r}, \hat{k})$ have a variation close to, but with a slightly smaller bias than the ones based on (r^*, k^*). Thus the selected $\hat{r}$ and $\hat{k}$ performed well. The smooth method of Storey and Tibshirani (2003) has a larger variation, and a larger bias, than the proposed method. Except for differences in the biases, the proposed method has variation similar to the convex decreasing density method.

Impact of Dependence: In this simulation, n =3,000, $m = 10$ and $B = 500$. First, two-sample gene expression data $\{x_{ij}, y_{ij} : i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, N\}$, $N = 10$, were generated in three different ways. For each gene i, the null hypothesis was $H_i : \mu_x = \mu_y$. Then B sets of p-values p_i of two-sample t-test with the same variances were calculated based on $x_{i1}, \ldots, x_{iN}$ and $y_{i1}, \ldots, y_{iN}$.

(1) *Independence:* $x_{ij} = \mu_{ij}$, $y_{ij} = \mu'_{ij} + 3I\{i \le (1 - \pi_0)n\}$, with μ_{ij}'s and μ'_{ij}'s i.i.d. $N(0, 1)$;

(2) *Dependence Case* 1: $x_{ij} = (-1)^i \mu_{vj} + \varepsilon_{ij}$, $y_{ij} = (-1)^i \mu'_{vj} + \varepsilon'_{ij} + 2I\{i \le (1-\pi_0)n\}$, $j = 1, \ldots, N$, $i = (v-1)m+1, \ldots, (v-1)m+m$, $v = 1, \ldots, n/m$, where ε_{ij}'s and ε'_{ij}'s are i.i.d. $N(0, 0.04^2)$ and, for each v, μ_{vj}'s and μ'_{vj}'s were i.i.d. $N(0, 1)$. In this case, the correlation between p-values for each pairs of genes in a group of m was 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: $x_{ij} = \mu_{ij} + \varepsilon_j$, $y_{ij} = \mu'_{ij} + \varepsilon'_j + 2I\{i \le (1 - \pi_0)n\}$, $j = 1, \ldots, N$, $i = 1, \ldots, n$, where ε_j's and ε'_j's were i.i.d. $N(0, 0.25^2)$. In this case, the correlation between p-values for each pairs of genes was about 0.0581 by simulation. This is the so-called "general dependence" of Storey (2002b).

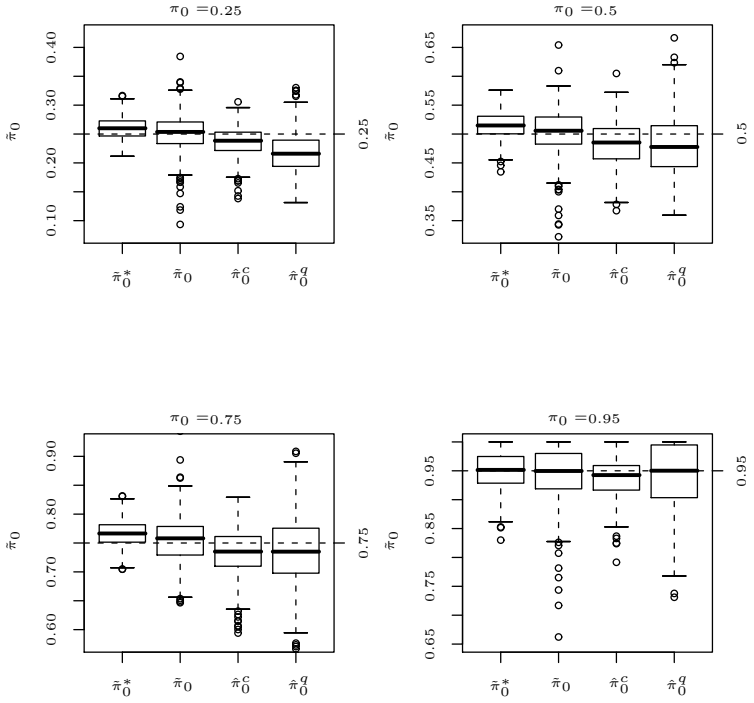

Figure 1. Simulation results for Beta distributed p-values. The true values of π_0 are $0.25, 0.50, 0.75$, and 0.95.

The simulated $\tilde{\pi}_0$'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 π_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 π_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.

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 types of tests. For each gene, a value of a test statistic is calculated based on sample

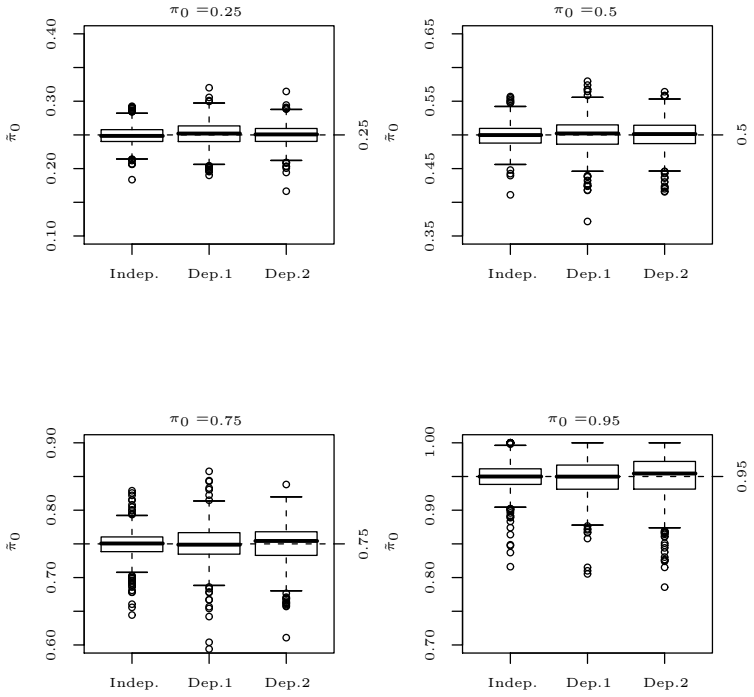

Figure 2. Simulation results for independent and dependent p-values. The true values of π_0 are $0.25, 0.50, 0.75$ and 0.95.

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.

Bias $\mathrm{E}(\tilde{\pi}_0) - \pi_0$					
π_0	0.05	0.25	0.50	0.75	0.95
Independence	0.00084	0.00095	0.00096	-0.00077	-0.00230
Dependence 1	-0.00057	0.00045	-0.00044	-0.00079	-0.00123
Dependence 2	-0.00075	-0.00115	-0.00114	-0.00068	-0.00109
Standard Deviation					
π_0	0.05	0.25	0.50	0.75	0.95
Independence	0.0083	0.0194	0.0252	0.0297	0.0303
Dependence 1	0.0075	0.0164	0.0230	0.0299	0.0347
Dependence 2	0.0071	0.0150	0.0182	0.0219	0.0237
Coverage Probability					
Independence	0.980	0.998	0.996	0.988	0.984
Dependence 1	0.968	0.976	0.960	0.930	0.916
Dependence 2	0.958	0.954	0.952	0.944	0.952

observations of the expression levels. Test statistics, such as p-values generated 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 7,129 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 =3,571 remaining genes after data preprocessing and filtering. The histogram of the p-values (not shown here) indicates that the mixture model (1.2) is valid, that the assumptions of the paper are not violated. Based on the expression data, correlation tests for the $n(n-1)/2$ pairs of genes, using a 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, data preprocessing and filtering are recommended, if possible, to have a smaller π_0 and thus a smaller variance of $\tilde{\pi}_0$. 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 cuto 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 4.944×10^{-3}.

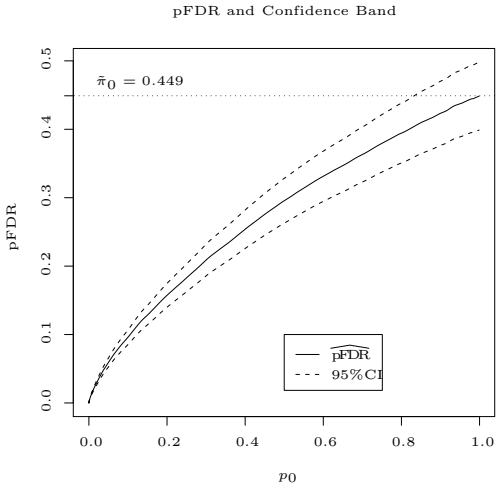

Figure 3. pFDR Estimation for the Golub et al. (1999) data.

Appendix

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

$$\tilde{\pi}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_{ki}, \tag{A.1}$$

$$Y_{ki} = k\sum_{j=0}^{k-1} \bar{b}(j,k,r) I\Big\{\frac{j}{k} < T_i \le \frac{j+1}{k}\Big\}, \quad \bar{b}(j,k,r) = \frac{1}{r}\sum_{l=1}^{r} B_{j,k-1}\Big(1-\frac{l}{k}\Big). \tag{A.2}$$

In the proof, the following results are useful. If $|g_j| \le M$, then for $i \ge 1$,

$$\sum_{j=0}^{k-1} \bar{b}^i(j,k,r) g_j = \mathcal{O}(1), \tag{A.3}$$

and for $v \ge i \ge 0$ and $v \ge 1$,

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

where $j^{[i]} = j(j-1)\cdots(j-i+1)$ if $j \ge i$; and is 0 otherwise. Define $j^{[0]} = 1$ for $j > 0$. Clearly, for each fixed k, $Y_{k1}, \ldots, Y_{kn}$ are identically distributed random variables with finite mean and variance given by, respectively,

$$\mu_k \equiv \mathrm{E}\,(Y_{ki}) = k\sum_{j=0}^{k-1} \bar{b}(j,k,r)\Delta_{kj}, \tag{A.5}$$

$$\sigma_k^2 \equiv \mathrm{Var}\,(Y_{ki}) = k^2\sum_{j=0}^{k-1} \bar{b}^2(j,k,r)\Delta_{kj} - \Big\{\mathrm{E}\,(Y_{ki})\Big\}^2, \tag{A.6}$$

where $\Delta_{kj} = F[(j+1)/k] - F(j/k)$. Since $f(1) = \pi_0$, $\mathrm{E}\,(\tilde{\pi}_0) = \mathrm{E}\,(Y_{k1})$ and $\sum_{j=0}^{k-1} \bar{b}(j,k,r) = 1$, the absolute bias is

$$B(r,k) \equiv |\mathrm{E}\,(\tilde{\pi}_0) - \pi_0| = |\mathrm{E}\,(Y_{k1}) - f(1)| = \Big|\sum_{j=0}^{k-1} \bar{b}(j,k,r)[k\Delta_{kj} - f(1)]\Big| .$$

Taylor expansions imply that the existence of $\xi_{0j} \in (j/k, (j+1)/k)$, $\xi_{1j} \in (j/k, j/(k-1))$, $\xi_{2j} \in (j/(k-1), 1-1/k)$ and $\xi_3 \in (1-1/k, 1)$ such that

$$k\Delta_{kj} = f\Big(\frac{j}{k}\Big) + \frac{1}{2k}f'(\xi_{0j}) = f\Big(\frac{j}{k-1}\Big) + \frac{1}{2k}f'(\xi_{0j}) - f'(\xi_{1j})\frac{j}{k(k-1)}$$

$$= f(1)+\frac{1}{2k}f'(\xi_{0j})-f'(\xi_{1j})\frac{j}{k(k-1)}-f'(\xi_{2j})\Big(1-\frac{1}{k}-\frac{j}{k-1}\Big)-f'(\xi_3)\frac{1}{k}. \tag{A.7}$$

Therefore

$$B(r,k) \le \sum_{i=0}^{3} R_{1i}(k,r), \tag{A.8}$$

where the $R_{1i}(k,r)$, $i=0,1,2,3$, are defined below. Since $tf'(t)$ is bounded,

$$\begin{aligned} kR_{1i}(k,r) &\equiv \Big(\frac{1}{2}\Big)^{1-i}\sum_{j=0}^{k-1}\bar{b}(j,k,r)|f'(\xi_{ij})|\Big(\frac{j}{k-1}\Big)^{i} \\ &\approx \Big(\frac{1}{2}\Big)^{1-i}\sum_{j=0}^{k-1}\bar{b}(j,k,r)\ f'\Big(\frac{j}{k-1}\Big)\ \Big(\frac{j}{k-1}\Big)^{i} = \mathcal{O}(1), \quad i=0,1. \end{aligned} \tag{A.9}$$

The Cauchy-Schwarz inequality implies

$$\begin{aligned} R_{12}^2(k,r) &\equiv \Big\{\sum_{j=0}^{k-1}\bar{b}(j,k,r)|f'(\xi_{2j})|\ 1-\frac{1}{k}-\frac{j}{k-1}\ \Big\}^2 \\ &\le \sum_{j=0}^{k-1}\bar{b}(j,k,r)\{f'(\xi_{2j})\}^2\sum_{j=0}^{k-1}\bar{b}(j,k,r)\Big(1-\frac{1}{k}-\frac{j}{k-1}\Big)^2. \end{aligned} \tag{A.10}$$

It follows from (A.4) that

$$\begin{aligned} \sum_{j=0}^{k-1}\bar{b}(j,k,r)\Big(1-\frac{1}{k}-\frac{j}{k-1}\Big)^2 &= \frac{1}{r}\sum_{l=1}^{r}\sum_{j=0}^{k-1}B_{j,k-1}\Big(1-\frac{l}{k}\Big)\Big(1-\frac{1}{k}-\frac{j}{k-1}\Big)^2 \\ &= \frac{1}{3k^2}\Big(r^2+2-\frac{r^2-1}{k-1}\Big). \end{aligned} \tag{A.11}$$

It follows from (A.3), (A.10) and (A.11) that

$$R_{12}(k,r) = \mathcal{O}(k^{-1}), \tag{A.12}$$

$$R_{13}(k,r) \equiv \sum_{j=0}^{k-1}\bar{b}(j,k,r)\frac{1}{k}|f'(\xi_3)| = \frac{1}{k}|f'(\xi_3)| \approx \frac{1}{k}\ f'\Big(1-\frac{1}{k}\Big)\ . \tag{A.13}$$

Combining (A.9) through (A.13) proves (2.2). For $u \neq v$,

$$\mathrm{Cov}\,(Y_{ku},Y_{kv}) = k^2\sum_{i=0}^{k-1}\sum_{j=0}^{k-1}\bar{b}(i,k,r)\bar{b}(j,k,r)\Big\{\Delta_{kij}^{(uv)}-\Delta_{ki}\Delta_{kj}\Big\}, \tag{A.14}$$

where

$$\Delta_{kij}^{(uv)} = \Pr\Big(\frac{i}{k} < T_u \le \frac{i+1}{k}, \frac{j}{k} < T_v \le \frac{j+1}{k}\Big)$$
$$= F_{uv}\Big(\frac{i+1}{k}, \frac{j+1}{k}\Big) - F_{uv}\Big(\frac{i}{k}, \frac{j+1}{k}\Big) - F_{uv}\Big(\frac{i+1}{k}, \frac{j}{k}\Big) + F_{uv}\Big(\frac{i}{k}, \frac{j}{k}\Big). \quad \text{(A.15)}$$

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 $\eta_j^{(uv)} \in (j/k, (j+1)/k)$,

$$k\Delta_{ki} = f(\tau_{ki}), \quad k^2\Delta_{kij}^{(uv)} = f_{uv}\Big(\zeta_i^{(uv)}, \eta_j^{(uv)}\Big). \quad \text{(A.16)}$$

Since f and f_{uv} are bounded, it follows from (A.14) and (A.16) that there exists a constant C such that for $u \neq v$

$$|\operatorname{Cov}(Y_{ku}, Y_{kv})| \le \max_{i,j}\Big\{f_{uv}(\zeta_i^{(uv)}, \eta_j^{(uv)}) + f(\tau_{ki})f(\tau_{kj})\Big\}\Big[\sum_{i=0}^{k-1}\bar{b}(i,k,r)\Big]^2 \le C,$$

$$\operatorname{Var}(\tilde{\pi}_0) - \frac{1}{n}\sigma_k^2 = \frac{1}{n^2}\sum_{u\neq v}\operatorname{Cov}(Y_{ku}, Y_{kv}) \le \frac{(\bar{n}-1)C}{n} \le \frac{m-1}{n}C. \quad \text{(A.17)}$$

From (A.3) it follows that

$$kR_{2i}(k,r) \equiv \sum_{j=0}^{k-1}\bar{b}^2(j,k,r)|f'(\xi_{ij})|\Big(\frac{j}{k-1}\Big)^i = \mathcal{O}(1), \quad i = 0, 1. \quad \text{(A.18)}$$

Another application of the Cauchy-Schwarz inequality gives

$$R_{22}^2(k,r) \equiv \Big\{\sum_{j=0}^{k-1}\bar{b}^2(j,k,r)\ f'(\xi_{2j})\ \ 1 - \frac{1}{k} - \frac{j}{k-1}\ \Big\}^2$$
$$\le \sum_{j=0}^{k-1}\bar{b}^3(j,k,r)\ f'(\xi_{2j})\ ^2\sum_{j=0}^{k-1}\bar{b}(j,k,r)\Big(1 - \frac{1}{k} - \frac{j}{k-1}\Big)^2$$
$$= \frac{1}{3k^2}\Big\{\sum_{j=0}^{k-1}\bar{b}^3(j,k,r)|f'(\xi_{2j})|^2\Big\}\Big\{r^2 + 2 - \frac{r^2-1}{k-1}\Big\} = \mathcal{O}(r^2k^{-2}), \quad \text{(A.19)}$$

$$R_{23}(k,r) \equiv \frac{1}{k}|f'(\xi_3)|h_k(r) = \mathcal{O}(k^{-1}). \quad \text{(A.20)}$$

From these, it follows that

$$k\sum_{j=0}^{k-1}\bar{b}^2(j,k,r)\Delta_{kj} - f(1)h_k(r) = \mathcal{O}(k^{-1}). \quad \text{(A.21)}$$

This, (2.5), (A.6), and (A.17) imply

$$\frac{\mathrm{Var}\,(Y_{ki})}{kh_k(r)} = \pi_0 - \frac{\mu_k}{kh_k(r)} + \mathcal{O}(k^{-2}) = \pi_0 + \mathcal{O}(k^{-1}),$$

$$\frac{n\mathrm{Var}\,(\tilde{\pi}_0)}{kh_k(r)} = \frac{\mathrm{Var}\,(Y_{ki})}{kh_k(r)} + \mathcal{O}(k^{-1}).$$

Consequently

$$\lim_{k,n\to\infty} \frac{n\mathrm{Var}\,(\tilde{\pi}_0)}{kh_k(r)} = \lim_{k\to\infty} \frac{\mathrm{Var}\,(Y_{ki})}{kh_k(r)} = \pi_0, \tag{A.22}$$

$$\mathrm{E}\,(\tilde{\pi}_0 - \pi_0)^2 = \mathcal{O}(k^{-2}) + \mathcal{O}\Big(\frac{k}{n}\Big).$$

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

$$I_0(2l)e^{-2l} = \sum_{j=0}^{\infty} \Big(\frac{l^j}{j!}e^{-l}\Big)^2 = \Pr\Big\{ \sum_{i=1}^{l} X_i - \sum_{i=1}^{l} Y_i = 0 \Big\}. \tag{A.23}$$

The Local Limit Theorem (see Petrov (1975, pp.187-188)) ensures that

$$\lim_{l\to\infty} \sqrt{l}\Pr\Big\{ \sum_{i=1}^{l} (X_i - Y_i) = 0 \Big\} = \frac{1}{2\sqrt{\pi}}.$$

From this it follows that there are constants $0 < C_1 < C_2$ such that

$$\frac{C_1}{\sqrt{l}} \le I_0(2l)e^{-2l} \le \frac{C_2}{\sqrt{l}}, \quad \text{for } l \ge 1, \tag{A.24}$$

$$I_0(2\sqrt{ij})e^{-i-j} = I_0(2\sqrt{ij})e^{-2\sqrt{ij}}e^{-(\sqrt{i}-\sqrt{j})^2} \begin{cases} \ge C_1(ij)^{-\frac{1}{4}}e^{-(\sqrt{i}-\sqrt{j})^2}; \\ \le C_2(ij)^{-\frac{1}{4}}e^{-(\sqrt{i}-\sqrt{j})^2}. \end{cases} \tag{A.25}$$

Combining (2.7), (A.23)– (A.25), one obtains

$$\begin{aligned} h(r) &\le C_2\Big\{ \frac{1}{r^2}\sum_{l=1}^{r} l^{-\frac{1}{2}} + \frac{2}{r^2} \sum_{1\le i<j\le r} e^{-(\sqrt{i}-\sqrt{j})^2}(ij)^{-\frac{1}{4}} \Big\} \\ &\le \frac{2C_2}{r^{\frac{3}{2}}}\Big\{ \int_0^1 t^{-\frac{1}{2}}dt + \frac{4}{\sqrt{r}}\int_0^{\sqrt{r}} dv \int_0^v \sqrt{uv}e^{-(u-v)^2}du \Big\} \le C_2' r^{-\frac{3}{2}}. \end{aligned}$$

Similarly, $h(r) \ge C_1' r^{-3/2}$. The proof of Theorem 2.1 is complete.

Proof of Theorem 2.2. Let

$$\xi_i = \frac{Y_{ki} - \mathrm{E}\,(\tilde{\pi}_0)}{n\sqrt{\mathrm{Var}\,(\tilde{\pi}_0)}}, \quad i = 1, 2, \ldots, n.$$

Then ξ_i has mean zero and $W = \sum_{i=1}^n \xi_i$ has variance one. By Assumption 1, $\xi_1, \ldots, \xi_n$ are also LD1 random variables. For each i, let η_i be the sum of all the random variables $\xi_{i_1}, \ldots, \xi_{i_{n_i}}$ that are not independent of ξ_i. By Theorem 3.4 of Chen and Shao (2005), we have

$$\sup_x |\Pr(W \le x) - \Phi(x)| \le 2\delta^{\frac{1}{2}}, \tag{A.26}$$

where

$$\begin{aligned}
\delta &= 4\mathrm{E}\ \sum_{i=1}^n \{\xi_i\eta_i - \mathrm{E}\,(\xi_i\eta_i)\}\ + \sum_{i=1}^n \mathrm{E}\,(|\xi_i\eta_i^2|) \equiv \delta_1 + \delta_2,\\
\delta_1 &\equiv 4\mathrm{E}\ \sum_{i=1}^n \{\xi_i\eta_i - \mathrm{E}\,(\xi_i\eta_i)\}\ = 4\mathrm{E}\ \sum_{i=1}^n \sum_{j=1}^{n_i} \{\xi_i\xi_{i_j} - \mathrm{E}\,(\xi_i\xi_{i_j})\}\\
&\le \frac{4}{n^2\mathrm{Var}\,(\tilde{\pi}_0)}\Big\{\mathrm{E}\ \sum_{i=1}^n \sum_{j=1}^{n_i} [Y_{ki}Y_{ki_j} - E(Y_{ki}Y_{ki_j})]\\
&+ \mathrm{E}\ \sum_{i=1}^n \sum_{j=1}^{n_i} \mathrm{E}\,(Y_{ki})[Y_{ki_j} - \mathrm{E}\,(Y_{ki_j})]\ + \mathrm{E}\ \sum_{i=1}^n \sum_{j=1}^{n_i} \mathrm{E}\,(Y_{ki_j})[Y_{ki} - \mathrm{E}\,(Y_{ki})]\ \Big\}\\
&\equiv \delta_{11} + \delta_{12} + \delta_{13}.
\end{aligned}$$

It is easy to see that there exists C_3 such that

$$\begin{aligned}
&\mathrm{E}\left[|Y_{ki}Y_{kj} - E(Y_{ki}Y_{kj})|\right]\\
&\le k^2 \sum_{u=0}^{k-1}\sum_{v=0}^{k-1} \bar{b}(u,k,r)\bar{b}(v,k,r)\mathrm{E}\ I\Big\{\frac{u}{k} < T_i \le \frac{u+1}{k}; \frac{v}{k} < T_j \le \frac{v+1}{k}\Big\} - \Delta_{kuv}^{(ij)}\\
&= 2k^2 \sum_{u=0}^{k-1}\sum_{v=0}^{k-1} \bar{b}(u,k,r)\bar{b}(v,k,r)\Delta_{kuv}^{(ij)}(1 - \Delta_{kuv}^{(ij)}) \le C_3.
\end{aligned}$$

Therefore

$$\begin{aligned}
\delta_{11} &\equiv \frac{4}{n^2\mathrm{Var}\,(\tilde{\pi}_0)}\mathrm{E}\ \sum_{i=1}^n \sum_{j=1}^{n_i} \left[Y_{ki}Y_{ki_j} - E(Y_{ki}Y_{ki_j})\right]\\
&\le \frac{4C_3 m}{n\mathrm{Var}\,(\tilde{\pi}_0)} = \mathcal{O}\Big(\frac{1}{kh_k(r)}\Big) = \mathcal{O}\Big(\frac{r^{\frac{3}{2}}}{k}\Big).
\end{aligned} \tag{A.27}$$

Similarly there exists C_4 such that, for any j,

$$\mathrm{E}\,[|Y_{kj} - E(Y_{kj})|] \le k\sum_{u=0}^{k-1} \bar{b}(u,k,r)\mathrm{E}\ I\Big\{\frac{u}{k} < T_j \le \frac{u+1}{k}\Big\} - \Delta_{ku}$$

$$= 2k \sum_{u=0}^{k-1} \bar{b}(u,k,r)\Delta_{ku}(1-\Delta_{ku}) \le C_4.$$

Thus

$$\delta_{12} = \frac{4}{n^2 \mathrm{Var}\,(\tilde{\pi}_0)} \mathrm{E} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \mathrm{E}\,(Y_{ki}) \Big[Y_{ki_j} - \mathrm{E}\,(Y_{ki_j}) \Big]$$

$$\le \frac{4C_4 m \mu_k}{n \mathrm{Var}\,(\tilde{\pi}_0)} = \mathcal{O}\Big(\frac{1}{k h_k(r)}\Big) = \mathcal{O}\Big(\frac{r^{\frac{3}{2}}}{k}\Big). \tag{A.28}$$

Similarly $\delta_{13} = \mathcal{O}\big(r^{3/2}/k\big)$. So $\delta_1 = \mathcal{O}\big(r^{3/2}/k\big)$,

$$\delta_2 \equiv \sum_{i=1}^{n} \mathrm{E}\,(|\xi_i \eta_i^2|) \le \frac{1}{n^3 \mathrm{Var}^{\frac{3}{2}}(\tilde{\pi}_0)} \sum_{i=1}^{n} n_i \sum_{j=1}^{n_i} \mathrm{E}\left[|Y_{ki} - \mu_k| |Y_{ki_j} - \mu_k|^2 \right]$$

$$\le \frac{k m^2 \,\mathrm{Var}\,(Y_{ki})}{n^2 \mathrm{Var}^{\frac{3}{2}}(\tilde{\pi}_0)} = r^{\frac{3}{4}} \mathcal{O}(\sqrt{k/n}).$$

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).

Acknowledgements

The authors would like to thank the two 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.

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A. and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. Data Anal.* **39**, 1-20.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Bernšteĭn, S. (1912). Démonstration du théorème de Weierstrass fondée sur le calcul des probabilities. *Comm. Soc. Math. Kharkov* **13**, 1-2.

Chen, L. H. Y. and Shao, Q.-M. (2005). Stein's method for normal approximation. In *An Introduction to Stein's Method* (Edited by A. D. Barbour and L. H. Y. Chen), 1-59. Number 4 in Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, Singapore University Press and World Scientific.

Efron, B. (2003). Robbins, empirical Bayes and microarrays. *Ann. Statist.* **31**, 366-378.

Efron, B. (2006). Correlation and large-scale simultaneous significance testing. Technical report, Standford University.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23**, 70-86.

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

Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test* **12**, 1-77.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B* **64**, 499-517.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.* **29**, 1264-1280.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Guan, Z., Wu, B. and Zhao, H. (2004). Model-based approach to FDR estimation. Research report 2004-016, Division of Biostatistics, University of Minnesota.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing* **3**, 135-146.

Langaas, M., Lindqvist, B. H. and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B* **67**, 555-572.

Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer-Verlag, New York.

Storey, J. D. (2002a). A direct approach to false discovery rates. *J. R. Statist. Soc. B* **64**, 479-498.

Storey, J. D. (2002b). False discovery rates: Theory and applications to DNA microarrays. Ph.D. thesis, Stanford University.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Ann. Statist.* **31**, 2013-2035.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B* **66**, 187-205.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Nat. Acad. Sci.* **100**, 9440-9445.

Tenbusch, A. (1994). Two-dimensional Bernstein polynomial density estimators. *Metrika* **41**, 233-253.

Vitale, R. A. (1975). Bernstein polynomial approach to density function estimation. In *Statistical Inference and Related Topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 2; dedicated to Z. W. Birnbaum)*, 87-99. Academic Press, New York.

Westfall, P. and Young, S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. Wiley, New York.

Wu, B., Guan, Z. and Zhao, H. (2006). Parametric and nonparametric FDR estimation revisited. *Biometrics* **62**, 735-744.

Yu, H. A. (1993). Glivenko-Cantelli lemma and the weak convergence for empirical processes of associated sequences. *Probab. Theory Relat. Fields* **95**, 357-370.

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

(Received January 2006; accepted October 2006)