JP Journal of Biostatistics
Volume 4, Number 2, 2010 Pages 181-201
Published Online: August 18, 2010
This paper is available online at http://pphmj.com/journals/jpb.htm

TESTING MULTIPLE HYPOTHESES USING POPULATION INFORMATION OF SAMPLES

MINGQI WU and FAMING LIANG*

Department of Statistics
Texas A & M University
College Station, TX 77843-3143, U. S. A.
e-mail: fliang@stat.tamu.edu

Abstract

Multiple hypothesis tests have been widely studied in the recent literature of statistics, however, most of the studies focus on how to control the false discovery rate for a given set of test scores or, equivalently, test *p*-values. Given the vast data involved in a multiple hypothesis test, it is natural to think about how to make use of population information of samples to improve the power of the test for each individual subject and thus to improve the power of the multiple hypothesis test. In this paper, we propose a nonparametric method for evaluation of test scores for each individual subject involved in a multiple hypothesis test. The method consists of two key steps, smoothing over neighboring subjects and density estimation over control samples, both of which allow for the use of population information of the subjects. The new method is tested on both the ChIP-chip data and the microarray data. The numerical results indicate that use of population information can significantly improve the power of multiple hypothesis tests.

1. Introduction

In biomedical study, many problems involve simultaneous tests of thousands, or

2010 Mathematics Subject Classification: 62G10, 62P10.

Keywords and phrases: ChIP-chip, density estimation, false discovery rate, microarray, multiple hypothesis testing, smoothing.

*Corresponding author

Received December 27, 2009

even millions, of null hypotheses. For example, Cawley et al. [6] considered the problem of identification of human transcription factor binding sites via ChIP-chip experiments, where more than 300,000 hypotheses were evaluated simultaneously; and Gottardo et al. [11] considered the problem of detection of differentially expressed genes under HIV-infected and non-infected conditions using microarrays, where 7680 hypotheses (genes) were tested simultaneously. How to effectively use the vast data in multiple hypothesis tests poses a great challenge for statisticians. The conventional multiple hypothesis testing procedure consists of the following typical steps:

- Sample collection: Let $X_1, ..., X_{r_1}$ denote the samples collected under the control condition, and let $Y_1, ..., Y_{r_2}$ denote the samples collected under the treatment condition. In a microarray experiment, for example, $X_i = (x_{1i}, ..., x_{ni})'$ is a vector of gene expression levels measured on array *i* and *n* is the number of genes involved in the experiment. This is the same for *Y*. Henceforth, *X*'s and *Y*'s are called the *control and treatment samples*, respectively; and $(x_{k,1}, ..., x_{k,r_1})$ and $(y_{k,1}, ..., y_{k,r_2})$ are called the *control and treatment samples* of subject *k*, respectively.

- Test score or *p*-value evaluation: This is hypothesis dependent. For example, to test the mean difference between the control and treatment samples, the two-sample Welch *t*-statistic (Welch [30]) is often used under the assumption that the experimental samples of each subject are mutually independent and normally distributed. The *p*-value of subject *k* can then be calculated accordingly.

- Significant subject identification: This can be done with various criteria, e.g., the per-comparison error rate, the family-wise error rate (Dudoit et al. [8]), and the false discovery rate (FDR) (Benjamini and Hochberg [1], Efron [9]).

Although the above procedure has succeeded in many applications, a drawback of the procedure is that the power of each individual test is low. This is because the sample replicates r_1 and r_2 are usually small and each individual test only makes use of sample information from the subject that it is testing. Given the vast data involved in a multiple hypothesis test, it is natural to think about how to make effective use of population information of samples to improve the power of the test for each individual subject and thus to improve the power of the multiple hypothesis test.

In this paper, we propose a nonparametric method for evaluation of test scores for each individual subject. The method consists of two key steps, smoothing over neighboring subjects and density estimation over control samples, both of which allow for the use of population information of the subjects. The new method is tested on both the ChIP-chip data and gene expression data. The numerical results indicate that use of population information can significantly improve the power of multiple hypothesis tests. In other words, the proposed method can significantly reduce the number of duplicates of the routine ChIP-chip and microarray experiments and thus the experimental cost, while maintaining the same level of statistical power in the analysis.

The remainder of this paper is organized as follows: In Section 2, we describe our new method how test scores are evaluated for individual subjects using population information of samples. In Sections 3 and 4, we apply, respectively, the new method to the ChIP-chip data and gene expression data along with comparisons with some existing multiple hypothesis testing methods. In Section 5, we conclude the paper with a brief justification for the rationale of the new method.

2. Testing Multiple Hypotheses using Population Information of Samples

In this section, we first describe a nonparametric method for evaluation of test scores for each individual subject under the assumption that the control samples are homogeneous, and then describe a procedure on how to prepare homogeneous control samples. The control samples are said *homogeneous* if the samples are identically distributed over all subjects. Finally, we describe how to identify significant subjects using a stochastic approximation FDR method (Liang and Zhang [18]).

2.1. A test with homogeneous control samples

Suppose that r_1 control samples, $X_1, ..., X_{r_1}$, and r_2 treatment samples, $Y_1, ..., Y_{r_2}$, have been collected in the experiment, respectively; and that the control samples are homogeneous. Furthermore, suppose that we are interested in testing simultaneously the mean difference of the control and treatment samples of n subjects; that is, to test the hypotheses $H_{k0} : \mu_{x,k} = \mu_{y,k}$ versus $H_{k1} : \mu_{x,k} < \mu_{y,k}$ for $k = 1, ..., n$, where $\mu_{x,k}$ and $\mu_{y,k}$ denote, respectively, the means of the control and treatment samples of subject k. To make use of population information in the test, we propose the following procedure:

- Density estimation: Fit a base density for $X_i = (x_{1i}, x_{2i}, ..., x_{ni})'$ using a nonparametric density estimation method, e.g., estimating f with the kernel estimator of the form

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - x_{ji}}{h} \right), \tag{1}$$

where K is the kernel, and h is the bandwidth which can be selected using an empirical plug-in rule. As shown by Hall et al. [12], this estimator is consistent even for long-range dependent data. Its asymptotic expansion for the mean integrated squared error (MISE) agrees to the second order with that of independent data. Denote the CDF of the fitted density by F_{X_i}. Totally, r_1 base densities are obtained. The kernel CDF, F_X, can be represented as

$$F_X = \frac{1}{r_1} \sum_{i=1}^{r_1} F_{X_i}, \tag{2}$$

by averaging over all base CDFs.

- Test score evaluation: Evaluate the p-value for each treatment sample of subject k using the kernel CDF F_X by

$$p_{k,j} = 1 - F_X(y_{k,j}), \quad j = 1, 2, ..., r_2, \quad k = 1, 2, ..., n, \tag{3}$$

and then evaluate the test score of subject k by

$$Z_k = \frac{1}{r_2} \sum_{j=1}^{r_2} \Phi^{-1}(1 - p_{k,j}), \quad k = 1, 2, ..., n, \tag{4}$$

which averages the test scores over all replicates.

Hereafter, this test will be called the *population-based test*. It has several significant advantages. First, it incorporates the population information of the control samples into the individual tests by basing the p-value evaluation on the fitted kernel distribution of the control samples. Second, it allows for the use of a single pair of control-treatment samples in multiple hypothesis tests; that is, both r_1 and r_2 can be as small as 1. Third, it is a nonparametric method which avoids the normality assumption for the samples. In our experience, the normality assumption is often violated by real biomedical data.

2.2. A general procedure for preparing homogeneous control samples

The key assumption of the population-based test is that the control samples are homogeneous, otherwise, the density estimation step is not sound. However, the raw control samples collected in biomedical study are usually not homogeneous. For example, this can be caused by the subject specific effect. Here, we propose a general procedure, which will transform the raw control samples to be homogeneous or approximately homogeneous. Note that our underlying assumption for the transformation is that the experimental samples follow a distribution in the location-scale family.

Let $\boldsymbol{a}_k = (x_{k,1}, ..., x_{k,r_1}, y_{k,1}, ..., y_{k,r_2})'$ represent the samples of subject k, $k = 1, ..., n,$ where the part $(x_{k,1}, ..., x_{k,r_1})$ denotes the control samples, and the part $(y_{k,1}, ..., y_{k,r_2})$ denotes the treatment samples. The transformation procedure can be described as follows.

For $k = 1, 2, ..., n,$ do the following:

- Neighboring subject identification: Calculate the distance between subject k and all other subjects using

$$d(\boldsymbol{a}_k, \boldsymbol{a}_s) = \| \boldsymbol{a}_k - \boldsymbol{a}_s \|, \quad \text{for} \quad s = 1, ..., n, \tag{5}$$

where $\| \boldsymbol{z} \|$ denotes the Euclidean norm of the vector $\boldsymbol{z}$. Identify l nearest subjects in terms of distance $d(\cdot, \cdot)$. The l subjects are called the neighboring subjects of subject k. The l is a predetermined number, depending on the problem under study. How to choose l will be discussed later; its effect will be measured in our simulated microarray data example.

- Smoothing: Smooth the samples of subject k by weightedly averaging the samples of the neighboring subjects. The method of weight assignment is also problem dependent. Generally speaking, the magnitude of the weight assigned to a neighboring subject should be reversely correlated to its distance to subject k.

- Standardization: Let $\boldsymbol{a}_k^* = (x_{k,1}^*, ..., x_{k,r_1}^*, y_{k,1}^*, ..., y_{k,r_2}^*)'$, $k = 1, 2, ..., n,$ denote the smoothed samples of subject k. Let $v_{x^*,k}$ and $v_{y^*,k}$ denote the James-Stein shrinkage variance (Opgen-Rhein and Strimmer [20]) of the smoothed control and smoothed treatment samples, respectively. Thus,

$$v_{x^*,k} = \lambda s^2_{x^*,median} + (1-\lambda) s^2_{x^*,k}, \tag{6}$$

where $s^2_{x^*,k}$ is the sample variance of $(x^*_{k,1}, ..., x^*_{k,r_1})$, $s^2_{x^*,median}$ is the median of $s^2_{x^*,k}$'s, and λ is the pooling parameter defined by

$$\lambda = \min\left(1, \frac{\sum_{k=1}^{n} \widehat{Var}(s^2_{x^*,k})}{\sum_{k=1}^{n} (s^2_{x^*,k} - s^2_{x^*,median})^2}\right), \tag{7}$$

where

$$\widehat{Var}(s^2_{x^*,k}) = \frac{r_1}{(r_1-1)^3} \sum_{i=1}^{r_1} (w_{ki} - \bar{w}_k)^2,\ w_{ki} = (x^*_{k,i} - \bar{x}^*_k)^2,\ \bar{w}_k = \frac{1}{r_1} \sum_{i=1}^{r_1} w_{ki},$$

and $\bar{x}^*_k = \frac{1}{r_1} \sum_{i=1}^{r_1} x^*_{k,i}$. The $v_{y^*,k}$ can be defined similarly. Given $v_{x^*,k}$ and $v_{y^*,k}$, we estimated the pooled variance of the samples of subject *k* by

$$\hat{\sigma}^2_k = \frac{(r_1-1)v_{x^*,k} + (r_2-1)v_{y^*,k}}{r_1 + r_2 - 2}. \tag{8}$$

Note that setting $\lambda = 0$, $\hat{\sigma}^2_k$ is reduced to the conventional pooled variance estimator for two samples.

Then, under the null hypothesis $H_{k0}: \mu_{x,k} = \mu_{y,k}$, we standardize the control and treatment samples of subject *k* by

$$\tilde{x}_{k,i} = \frac{x^*_{k,i} - \bar{x}^*_k}{\hat{\sigma}_k}, \quad \tilde{y}_{k,j} = \frac{y^*_{k,j} - \bar{x}^*_k}{\hat{\sigma}_k}, \tag{9}$$

for $i = 1, ..., r_1$ and $j = 1, ..., r_2$.

It is clear that the samples $\tilde{x}_{k,i}$'s are identically distributed under the mild assumption that the original samples $x_{k,i}$'s follow a distribution in the location-scale family. Thus, the transformed control samples are homogeneous, and the density estimation described in Subsection 2.1 is applicable. This approach is similar to Song and Hart's cluster-based density estimate [25]. Note that the transformation procedure has been designed to incorporate information from other subjects. This reflects in two steps, smoothing over neighboring subjects and calculation of James-Stein shrinkage variance. As argued at the end of the paper, smoothing over

neighboring subjects reduces effectively the variation of the experimental samples, while causing only negligible bias to the mean of the samples as long as *l*, the size of neighboring subject set, is reasonable.

In this subsection, we only outline the idea how to prepare homogeneous control samples by using population information of the samples. In practice, many detailed steps, such as determination of neighboring subjects and smoothing weight assignment, will depend on the problem under study. In Sections 3 and 4, we will give details on how the idea works for the ChIP-chip data and the gene expression data.

2.3. FDR control

Given the test scores, a multiple hypothesis testing procedure is still needed for identification of significant subjects. Here, we adopted the stochastic approximation-based FDR control method developed by Liang and Zhang [18], which, hereafter, will be abbreviated as the SA-FDR method. The SA-FDR method falls into the class of empirical Bayes methods (Efron [9]). Like other methods in this class, it works by fitting the test scores with a two-component mixture model

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z), \tag{10}$$

where π_0 is the prior probability that a null hypothesis is true, f_0 is the empirical null distribution, f_1 is the alternative distribution, and f_0 is stochastically smaller than f_1. Given the estimators of π_0 and f_0, the positive FDR (Storey [28]) of a rejection rule $\Lambda = \{Z_i \geq z_0\}$ can be estimated by

$$\widehat{\mathrm{Fdr}}(\Lambda) = \frac{N\hat{\pi}_0 [1 - \hat{F}_0(z_0)]}{\#\{z_i : z_i \geq z_0\}}, \tag{11}$$

where $\#\{z_i : z_i \geq z_0\}$ denotes the number of subjects with test scores greater than z_0, $\hat{\pi}_0$ denotes the estimator of π_0, and $\hat{F}_0$ denotes the CDF estimator of f_0. Note that $\widehat{\mathrm{Fdr}}(\Lambda)$ can be intuitively interpreted as the expected proportion of null subjects, i.e., the subjects with the null hypotheses being true, among those with the test score greater than z_0. Following the suggestion by Storey [26], the *q*-value defined below

$$q(z) \equiv \inf_{\{\Lambda: z \in \Lambda\}} \mathrm{Fdr}(\Lambda), \tag{12}$$

is used in this paper as a reference quantity for the decision of multiple hypothesis testing.

In Liang and Zhang [18], π_0 and f_0 are estimated using a two-step procedure:

- Fit the distribution of the test scores with a mixture of exponential power distributions using the stochastic approximation method (Robbins and Monro [22], Benveniste et al. [3]).

- Clustering the components of the mixture exponential power distributions into two clusters, which correspond to f_0 and f_1 of the mixture (10), respectively, according to the mutual distance between the components.

Liang and Zhang [18] showed theoretically that the method is valid under general dependence between test scores. We note that for the population-based test proposed in this paper, the use of the SA-FDR method is not essential. Any other multiple comparison methods, e.g., the methods developed by Benjamini and Yekutieli [2], Storey et al. [28], and Efron [9], can be equally used here. To use the method proposed by Benjamini and Yekutieli [2] and Storey et al. [28], we may need to transform the test scores to *p*-values via the transformation $P = 1 - \Phi^{-1}(Z)$.

3. Application of the Population-based Test to ChIP-chip Data Analysis

In this section, we applied the population-based test to ChIP-chip data for the purpose of identification of transcription factors binding sites (TFBS). The performance of our method is first assessed on a real dataset, and then assessed on some simulated datasets.

3.1. p53 data

The dataset, we studied here, was generated by Cawley et al. [6], whose experiment mapped the binding sites of four human transcription factors Sp1, cMyc, p53-FL, and p53-DO1 on chromosomes 21 and 22. The chromosomes were spanned over three chips A, B and C. All experiments were done under three conditions: IP, control GST and control input. For each transcription factor, under each experiment condition, 6 samples (2 biological replicates $\times 3$ technical replicates) were obtained. For the testing purpose, p53-FL data on chips A, B and C, under IP and control input conditions, were analyzed in this paper. The raw data is available at http://transcriptome.affymetrix.com/publication/tfbs.

For comparison, the raw data were pre-processed as in Cawley et al. [6]. We first filtered out the local repeats, and then normalized the data using the quantile-

normalization method (Bolstad et al. [4]). After normalization, the data were rescaled to have a median feature intensity of 1000, log-transformed, and then processed as prescribed in Subsection 2.2. For the ChIP-chip data, the neighbor identification step can be skipped, because, by the nature of the data, the probes have been self-clustered into bound and non-bound regions. For the smoothing step, the Gaussian weighted moving average method was applied as in Zheng et al. [32]. A window with size $1000bp(\pm 500bp)$ was moving along the genome. The intensity of the probe in the center of the window is updated by

$$\boldsymbol{a}_k^* = \sum_{i \in window} w_i \boldsymbol{a}_i \Big/ \sum_{i \in window} w_i, \quad w_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{d_{k,i}^2}{2\sigma^2}\right), \tag{13}$$

where $\boldsymbol{a}_i$ denotes the intensity values of probe *i* measured in the experiment, $d_{k,i}$ is the genomic distance between the central positions of probe *i* and probe *k*, and the standard deviation is set to be one fourth of the window size, $\sigma = 250bp$. Note that the probe-specific effect has been removed by sample centralization in the standardization step (9). Hence, the control samples are, at least approximately, homogeneous after pre-processing.

Following Keles et al. [15], we define a scan-statistic, which is a moving average of the test scores resultant from the population-based method, i.e.,

$$Z_k^s = \frac{1}{2w+1} \sum_{i=k-w}^{k+w} Z_i, \tag{14}$$

where $w = 5$ is the half moving window size, which is the same as that used in Ji and Wong [13], Keles et al. [15] and Gottardo et al. [10]. By doing so, information from neighboring probes were further borrowed for identification of bound regions.

In Cawley et al. [6], a cutoff of 10^{-5} was used for the *p*-values resultant from the Wilcoxon rank sum test, and this led to 36, 353 and 423 probes being identified as "significant" probes on chips A, B and C, respectively. For comparison, we set the cutoff numbers to 36, 353 and 423 for the test scores on chips A, B and C, respectively. Following the approach taken by Cawley et al. [6], the regions having less than 3 probes or 100 bps were considered to be spurious and removed, and the regions separated by 500 bps or less were merged together to form a predicted bound region. The results are summarized in Table 1. The Wilcoxon rank sum test

identified 9 out of 14 experimentally validated bound regions, while the population-based test identified 12 out of the 14 validated bound regions. For a further comparison, we relaxed the cutoff number and counted the total number of "significant" probes needed to cover all the 14 validated bound regions. For the Wilcoxon rank sum test, it needs to increase the total number of "significant" probes to 7292; while for the population-based method, it only requires 1241 "significant" ones. The population-based test outperforms significantly the Wilcoxon rank sum method for this example.

Table 1. Computational results for the p53-FL data. *V*: the number of bound regions that have been experimentally validated and identified by the method; *a*: the cutoff number specified by Cawley et al. [6]; *b*: the number of bound regions that have been experimentally validated on the chip; and τ^*: the number of "significant" probes needed to cover all experimentally validated bound regions

S	Chip A		Chip B		Chip C	
Method	$V(36^a, 2^b)$	τ^*	$V(353^a, 2^b)$	τ^*	$V(423^a, 10^b)$	τ^*
Wilcoxon	2	29	1	862	6	6401
Population-based	2	34	2	71	8	1136

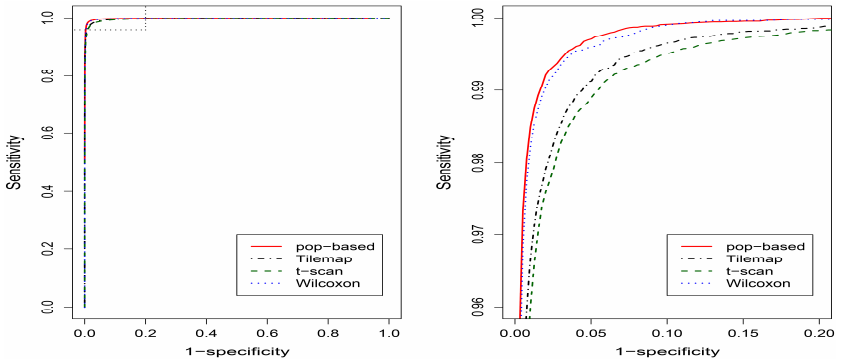
3.2. Simulated data

To have a careful assessment of the performance of the population-based test on ChIP-chip data, we simulated 20 datasets based on the Sp1 data generated by Cawley et al. [6]. Each dataset consists of 200,000 probes, two conditions (IP and control input), and six replicates under each condition. We extract the first 200,000 genomic positions of the Sp1 data as the probe genomic coordinates in the simulations. Each dataset consisted of 996 bound probes, which form 50 bound regions. As in Gottardo et al. [10], the bound regions were assumed to describe a peak with the intensity function given by $A\exp\{-4(g_i - C)^2/B^2\}$, where *A* is the amplitude of the peak, *B* controls the width of the peak, *C* represents the center of the peak, and g_i is the genomic position of probe *i*. We also followed Gottardo et al. [10] to generate the centers of the bound regions randomly across the set of possible coordinates while imposing a separation of at least 3000 bps between peaks; and to generate the values of parameter *B* uniformly between 600 and 1000 bps. The values of parameter *A* were generated uniformly between 3 and 5. The variance of the probe intensity was estimated from the Sp1 data.

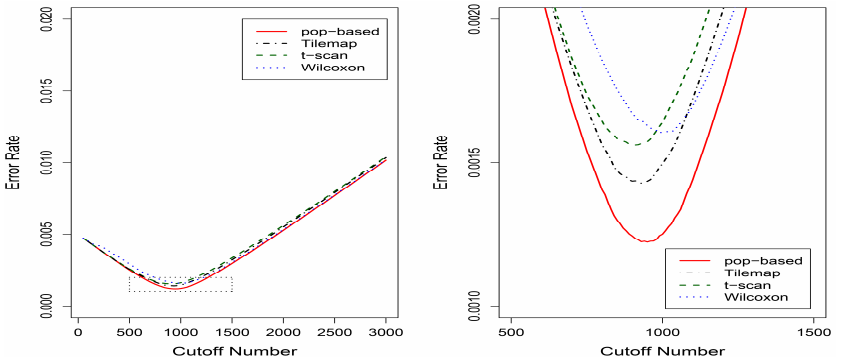
For comparison, four different methods were applied to the 20 simulated datasets, including the Wilcoxon rank sum test (Cawley et al. [6]), *t*-scan test (Keles et al. [15]), Tilemap (Ji and Wong [13]), and population-based test. The results are summarized in Table 2. For testing purpose, we tried different cutoff numbers for the probes. At each of the cutoff numbers, the methods were compared in three criteria, the number of false negative bound regions (i.e., the number of bound regions not identified by the method), the number of false positive bound regions (i.e., the number of falsely identified bound regions), and the number of true positive probes (i.e., the number of correctly identified bound probes). The numerical results show that for this example, Tilemap works better than the *t*-scan test in terms of all three criteria, and the Wilcoxon method finds less false positive bound regions than Tilemap and *t*-scan. While, the population-based method outperforms other three methods in all three criteria.

Table 2. Computational results for the simulated datasets. At each cutoff number τ, the number of false negative bound regions (missed regions), the number of false positive bound regions (extra regions), and the number of true positive probes (matched probes) with their standard deviations (the numbers in the parentheses) were calculated by averaging over the 20 datasets

		Methods			
τ	Criteria	Wilcoxon	*t*-scan	Tilemap	pop-based
800	Missed regions	8.1(.37)	4.9(.35)	4.5(.37)	**4.2(.29)**
	Extra regions	1.2(.29)	5.7(.69)	2.7(.31)	**0.9(.18)**
	Matched probes	712.3(2.62)	733.5(2.54)	741.8(1.85)	**757.4(2.15)**
900	Missed regions	5.6(.34)	3.3(.29)	3.3(.28)	**2.5(.21)**
	Extra regions	3.2(.52)	13.9(1.36)	7.9(.63)	**2.5(.29)**
	Matched probes	780.9(3.26)	791.9(3.63)	804.5(2.94)	**823.6(2.42)**
1000	Missed regions	3.2(.40)	2.5(.24)	2.6(.23)	**1.5(.15)**
	Extra regions	6.1(.66)	26.6(1.30)	19.8(1.16)	**6.4(.37)**
	Matched probes	837.3(3.82)	834.0(4.30)	848.5(3.32)	**872.9(2.79)**
1100	Missed regions	1.95(.36)	2.1(.23)	1.7(.21)	**1.0(.15)**
	Extra regions	15.6(.89)	44.7(1.92)	37.7(1.55)	**12.3(.60)**
	Matched probes	878.3(4.71)	861.3(4.07)	875.8(3.95)	**904.0(2.58)**
1200	Missed regions	1.5(.26)	1.6(.22)	1.4(.17)	**0.9(.17)**
	Extra regions	27.9(1.14)	67.8(2.48)	58.9(1.59)	**21.1(.75)**
	Matched probes	905.0(4.47)	879.4(4.24)	894.8(3.49)	**923.0(2.55)**
1500	Missed regions	0.7(.16)	1.2(.17)	1.1(.20)	**0.5(.11)**
	Extra regions	84.6(2.27)	144.1(3.14)	141.6(2.69)	**48.9(0.99)**
	Matched probes	943.7(4.22)	912.0(4.00)	925.6(3.24)	**950.5(2.26)**



(a) ROC curves.



(b) Error rate.

Figure 1. Averaged ROC curves and error rate curves (over 20 datasets) for the population-based, Tilemap, *t*-scan and Wilcoxon methods. (a) the ROC curve; (b) the error rate curve. The right panel plot provides a closer view for the area enclosed by the dotted line and the axes in the left panel plot.

Later, we compared the receiving operating characteristic (ROC) curves (Bradley [5]) and the error rate curves for the four methods. The ROC curve shows the true positive discovery rate (sensitivity) against the false positive discovery rate (1-specificity) at the probe level, and the area under the curve (AUC) has been used as a summary measure of accuracy for multiple hypothesis tests. The error rate curve shows the proportion of incorrect probe calls, including both false positives and false negatives, against different cutoff values. The error rate has been used as a summary measure for the performance of a clustering method. The averaged ROC curve and error rate curve (over 20 datasets) are shown in Figure 1. In terms of AUCs, the four tests are ranked as the population-based test, Wilcoxon, Tilemap and *t*-scan, from the

best to the worst. This is a little different from our impression obtained from Table 2, where it seems that Tilemap and *t*-scan outperform the Wilcoxon method. An interpretation for the difference is that AUC emphasizes more on the false discovery rate. It is indeed that the Wilcoxon method consistently produced smaller numbers of false positive bound regions than do the Tilemap and *t*-scan methods for this example. Next, we examined the error rates of the four methods. Figure 1(b) indicates that all the four methods have an optimal cutoff number around 996, which is the number of true bound probes. It is remarkable that, among the four tests, the population-based test has consistently the lowest error rate at various cutoff values.

4. Application of Population-based Tests to Microarray Data Analysis

In this section, we considered the application of the population-based test to microarray data for identification of differentially expressed genes. First, we tested the new method on a simulated example, which was modified from some examples used in the literature. Next, we applied the new method to a real dataset which is typical in this area.

4.1. A simulated example

This example is modified from examples of Qiu et al. [21] and Liang et al. [17]. It consists of multiple simulated datasets. Let *n* denote the number of genes included in each dataset, and let *m* denote the number of differentially expressed genes. The datasets were generated in the following way.

First, generate an $n \times 6$ matrix and denote this matrix by $X = (x_{ij})$, $i = 1, ..., n$ and $j = 1, ..., 6$. The elements of this matrix are set as

$$x_{ij} = \begin{cases} [1.0]\mu_i + \sigma_i z_{ij}, & \text{if } i = 1, \\ \mu_i + \rho \dfrac{\sigma_i}{\sigma_{i-1}} (x_{i-1,\,j} - \mu_{i-1}) + \sigma_i \sqrt{1-\rho^2}\, z_{ij}, & \text{if } i = 2, ..., n, \end{cases} \tag{15}$$

where μ_i, σ_i and z_{ij} are drawn independently from the distributions

$$\mu_i \sim U(-0.5, 0.5), \quad \sigma_i \sim U(0.5, 1.5), \quad z_{ij} \sim N(0, 1). \tag{16}$$

It is not difficult to show that $\mathrm{Corr}(x_{i,\,j}, x_{i+1,\,j}) = \rho$ for $i = 1, ..., n-1$ and any *j*. In other words, there is constant correlation between the expression levels of adjacent genes.

Next, define

$$y_{ij} = \begin{cases} [1.0]x_{ij} + \mu, & \text{for } i = 1, ..., m, \quad j = 4, ..., 6, \\ x_{ij}, & \text{otherwise,} \end{cases} \tag{17}$$

where μ is a constant representing the mean expression level difference of the differentially expressed genes and nondifferentially expressed genes. For each dataset $Y = (y_{ij})$, $i = 1, ..., n$ and $j = 1, ..., 6$, generated in the above procedure, the first m rows model the differentially expressed genes, the first 3 columns represent the control samples, and the last 3 columns represent the treatment samples.

For comparison, we calculated the test scores using the following two methods:

- Score A: The population-based test with the three control and the three treatment samples. To prepare homogeneous control samples, the data are smoothed as follows,

$$\boldsymbol{a}_k^* = \sum_{i \in C_k} w_i \boldsymbol{a}_i, \quad w_i = \frac{\hat{\rho}_{i,k}^2}{\sum_{i \in C_k} \hat{\rho}_{i,k}^2}, \tag{18}$$

where $\boldsymbol{a}_i = (x_{i1}, x_{i2}, x_{i3}, y_{i1}, y_{i2}, y_{i3})'$, C_k denotes the set of neighboring genes of gene k, and $\hat{\rho}_{i,k}$ denotes the Pearson correlation coefficient between the expression levels of gene k and gene i. For this example, we set $|C_k|$, the size of C_k, to be 10 based on the belief that there are about 10 genes co-expressed with each gene in the dataset. As indicated by our numerical results presented below, the population-based test is rather robust to the size of C_k.

- Score B: Two-sample t-test with the three control and the three treatment samples.

Please note that the t-tests used in Score B is exact, as the data are generated from normal distributions. From this point of view, the comparison is a little unfair to Score A. In the following, we compare Score A and Score B, by looking at the power of the resulting multiple hypothesis tests.

4.1.1. Score A versus Score B

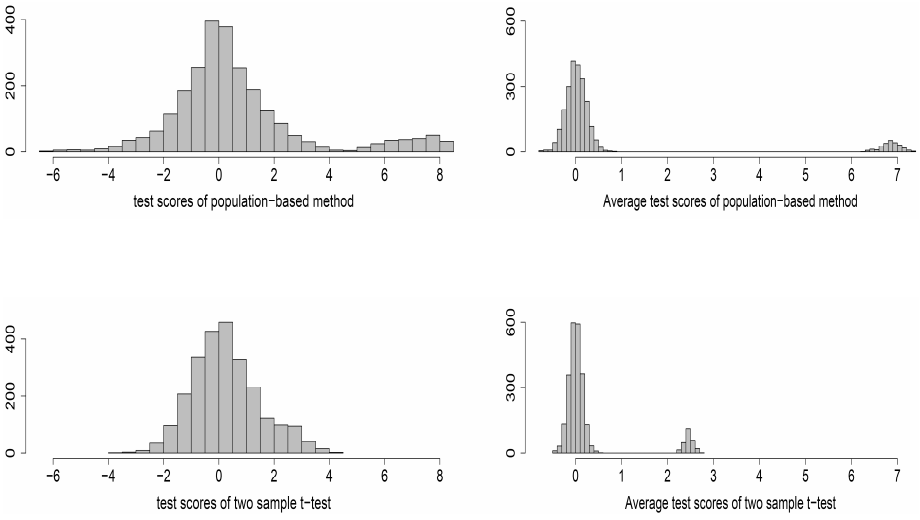
In this comparison, we fix $n = 2500$, $m = 250$, $\mu = 3$, $\rho = 0.3$, and $|C_k| = 10$, generated 50 different datasets in the above procedure and calculated Scores A

and B. Then the true FDRs (tFDRs) were calculated at different cutoff numbers. A cutoff number, τ, defines a classification criterion, classifying the τ genes with the highest test scores as "differentially" expressed genes. The computational results are summarized in Table 3. It shows that the tFDRs resultant from Score A is lower than those from Score B for all the chosen cutoff numbers. This implies that the multiple hypothesis test based on Score A can have a higher power than that based on Score B. In addition, we compare the histograms of Score A and Score B. Figure 2(a) shows the histograms of the test scores for one of the 50 datasets, which indicates that the distribution of the test scores produced by the population-based method has relatively heavier and longer tail in the significant region than that produced by the two-sample *t*-test. Hence, the differentially and non-differentially expressed genes can be better separated by Score A than by Score B. Furthermore, we examined the histograms of the averaged test scores, where the average is taken for each gene over the 50 datasets. As shown in Figure 2(b), the distance between differentially expressed genes and non-differentially ones is almost three times longer for the population-based method than for the two-sample *t*-test method. This supports again our claim that the differentially and non-differentially expressed genes can be better separated by Score A for this example.

Table 3. Computational results for the datasets generated with $n = 2500$, $m = 250$, $\mu = 3$, $\rho = 0.3$, and $|C_k| = 10$. The value of tFDR and its standard deviation (the number in the parentheses) were calculated by averaging over 50 datasets

	Cutoff number τ						
Method	50	100	200	250	300	400	500
Score A	.006(.002)	.005(.001)	.022(.002)	.084(.002)	.201(.002)	.387(.001)	.507(.001)
Score B	.051(.005)	.078(.004)	.165(.004)	.228(.003)	.294(.003)	.422(.002)	.520(.001)

In addition, comparisons between population-based test with three control and three treatment samples and two-sample *t*-test with four control and four treatment samples have also been carried out, with various choices of the parameters n, ρ, μ, and $|C_k|$ in terms of specificity and sensitivity of multiple hypothesis tests. The numerical results indicate that the population-based test can outperform the two-sample *t*-test in almost all scenarios for this example, which means to achieve the same or even higher testing power while maintaining the same level of specificity, the population-based test requires less than $3/4$ of the control and treatment samples than does the two-sample *t*-test. This implies that use of the population-based test can potentially lead to a great saving of experiment cost.



(a) (b)

Figure 2. Histograms of test score for the simulated data. Left panel: histograms of test scores of one dataset. Right Panel: histograms of the average test scores of 50 simulated datasets.

4.2. HIV data

This dataset includes $n = 7680$ genes. It concerns the difference of gene expression levels of uninfected cells and HIV-infected cells (Wout et al. [31]). As described by Gottardo et al. [11], the experiment was carried out on four different slides under the same RNA preparation. Each slide reported on the same set of 7680 genes. Among them, 12 known differentially expressed HIV-1 genes are included as positive controls. Dye-swapped hybridizations technique was used to compensate dye bias in this experiment. Two of the four slides were hybridized with the green dye (Cy3) for the control (uninfected) samples and the red dye (Cy5) for the treatment (HIV-infected) samples, the dyes were reversed on the other two slides. Totally, 4 control and 4 treatment samples were included in this dataset. The raw data was downloaded from http://wwwstatubcca/~raph/PublicFiles/.

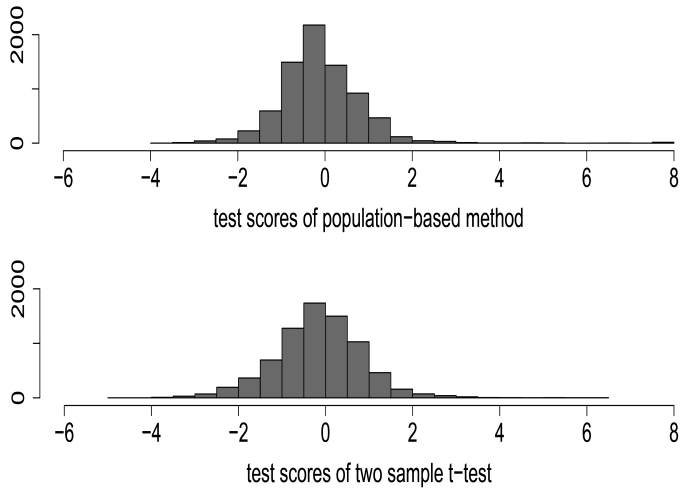


Figure 3. Test score of HIV dataset for population-based method and two-sample *t*-test.

The raw data were pre-processed as in Gottardo et al. [11]. They were first quantile-normalized (Bolstad et al. [4]), log-transformed, and then the mean of the log expression values was adjusted to zero for each chip. Afterwards, the two-sample *t*-test was applied to the pre-processed data. On the other hand, for the population-based test, the data were further processed as prescribed in Subsection 2.2. As for the simulated example, we set the size of C_k, the number of neighboring genes, to be 10, and smoothed the gene expression levels using (18).

Figure 3 displays the histograms of the test scores produced by the two methods. It is easy to see that the histogram produced by the population-based test has a relatively shorter left-tail and longer right-tail than that produced by the two-sample *t*-test. This difference implies that the population-based test can have a higher power than the two-sample *t*-test. The SA-FDR methods were applied to the test scores resultant from the two methods. By controlling the nominal FDR at 10%, only 16 differentially expressed genes were detected using the test scores produced by the two-sample *t*-test. This is too conservative, comparing to 33, 86 and 81 genes found by the software BRIDGE (Gottardo et al. [11]), the empirical Bayes gamma-gamma model (Newton et al. [19]) and the empirical Bayes lognormal-normal model (Kendziorski et al. [16]), respectively. Using the test scores resultant from the population-based test, 50 genes were identified as being differentially expressed. It is remarkable that, not only the built-in 12 positive control genes are covered by the 50

significant genes, but also their test scores are ranked among the top 13 test scores. This does not happen for the two-sample *t*-test. All the above evidences indicate the effectiveness of the population-based test for detecting differential expressed genes with microarray experiments.

5. Discussion

We have proposed a population-based method for evaluation of test scores for each individual subject involved in a multiple hypothesis test. The method consists of two key steps, smoothing over neighboring subjects and density estimation over control samples, both of which allow for the use of population information of the subjects. The new method is tested on both the ChIP-chip data and the gene expression data. The numerical results indicate that use of population information can significantly improve the power of multiple hypothesis tests. In other words, the proposed method can significantly reduce the number of duplicates of the routine ChIP-chip and microarray experiments and thus the experimental cost, while maintaining the same level of statistical power in the analysis.

The strength of the new method comes from two sources, smoothing over neighboring subjects and density estimation over control samples. Smoothing over neighboring subjects effectively reduces variation of the experimental samples, while causing only negligible bias to the mean of the samples as long as the size of neighboring subjects set is reasonable. As shown by our numerical examples, smoothing over neighboring subjects samples does improve the power of multiple hypothesis tests.

Nonparametric density estimation over control samples provides us a robust way of test scores evaluation, which relaxes the distribution assumption for the experimental samples from normality to a location-scale family. Moreover, it automatically accounts for the extremeness of a large size sample with its built-in mechanism. This is beyond the ability of the two-sample *t*-test and other tests based on the individual subject samples.

At last, we would like to mention that the idea of using population information of samples to improve the power of multiple hypothesis tests is not brand new. Smyth [24], Ji and Wong [13], Cui et al. [7], and Opgen-Rhein and Strimmer [20], can be viewed as early works in this research direction, although the idea of using population information was not stated there explicitly. In these works, the variance

of each individual subject samples was estimated by an empirical Bayesian approach or a shrinkage approach which make use of other subjects samples. Although the population information of samples is used very limitedly in these works, their numerical results do show that use of population information can improve the power of multiple hypothesis tests. In this paper, we made a full exploration of this idea by using the traditional nonparametric technique of kernel density estimation.

Acknowledgements

Liang's research was partially supported by grants from the National Science Foundation (DMS-0607755) and the award (KUS-C1-016-04) made by King Abdullah University of Science and Technology (KAUST). The authors would like to thank Dr. J. D. Hart for his helpful discussion on kernel density estimation for dependent data.

References

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

[2] Y. Benjamini and D. Yekutieli, On the control of false discovery rate in multiple testing under dependency, Ann. Statist. 29 (2001), 1165-1188.

[3] A. Benveniste, M. Métivier and P. Priouret, Adaptive Algorithms and Stochastic Approximations, Springer-Verlag, 1990.

[4] B. M. Bolstad, R. A. Irizarry, M. Astrand and T. P. Speed, A comparison of normalization methods for high density ologonucleotide array data based on variance and bias, Bioinformatics 19(2) (2003), 185-193.

[5] A. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (1997), 1145-1159.

[6] S. Cawley, S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl and T. R. Gingeras, Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs, Cell 116 (2004), 499-509.

[7] X. Cui, J. T. G. Hwang, J. Qiu, N. J. Blades and G. A. Churchill, Improved statistical tests for differential gene expression by shrinking variance components estimates, Biostatistics 6(1) (2005), 59-75.

[8] S. Dudoit, J. P. Shaffer and J. C. Boldrick, Multiple hypothesis testing in microarray experiments, Statist. Sci. 18 (2003), 71-103.

[9] B. Efron, Large-scale simultaneous hypothesis testing: the choice of a null hypothesis, J. Amer. Statist. Assoc. 99 (2004), 96-104.

[10] R. Gottardo, W. Li, W. E. Johnson and X. S. Liu, A flexible and powerful Bayesian hierarchical model for ChIP-chip experiments, Biometrics 64 (2008), 468-478.

[11] R. Gottardo, A. E. Raftery, K. Y. Yeung and R. E. Bumgarner, Bayesian robust inference for differential gene expression in microarrays with multiple samples, Biometrics 62 (2006), 10-18.

[12] P. Hall, S. N. Lahiri and Y. K. Truong, On bandwidth choice for density estimation with dependent data, Ann. Statist. 23 (1995), 2241-2263.

[13] H. Ji and W. H. Wong, TileMap: Create chromosomal map of tiling array hybridizations, Bioinformatics 21(18) (2005), 3629-3636.

[14] S. Keles, Mixture modeling for genome-wide localization of transcription factors, Biometrics 63 (2007), 10-21.

[15] S. Keles, M. J. Van Der Laan, S. Dudoit and S. E. Cawley, Multiple testing methods for ChIP-chip high density oligonucleotide array data, J. Comput. Biol. 13(3) (2006), 579-613.

[16] C. Kendziorski, M. Newton, H. Lan and M. N. Gould, On parameter empirical Bayes methods for comparing multiple groups using replicated gene expression profiles, Stat. Med. 22 (2003), 3899-3914.

[17] F. Liang, C. Liu and N. Wang, A robust sequential Bayesian method for identification of differentially expressed genes, Statist. Sinica 17 (2007), 571-597.

[18] F. Liang and J. Zhang, Estimating FDR under general dependence using stochastic approximation, Biometrika 95(4) (2008), 961-977.

[19] M. C. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Balattner and K. W. Tsui, On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, J. Comput. Biol. 8 (2001), 37-52.

[20] R. Opgen-Rhein and K. Strimmer, Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach, Stat. Appl. Genet. Mol. Biol. 6 (2007), Article 9, 10 pp.

[21] X. Qiu, L. Klebanov and A. Yakovlev, Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes, Stat. Appl. Genet. Mol. Biol. 4 (2005), article 34, 32 pp.

[22] H. Robbins and S. Monro, A stochastic approximation method, Ann. Math. Statist. 22 (1951), 400-407.

[23] D. Rubin, S. Dudoit and M. J. van der Laan, A method to increase the power of multiple testing procedures through sample splitting, 2006, Available at http://www.bepress.com/ucbbiostat/paper171

[24] G. K. Smyth, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (2004), article 3, 29 pp.

[25] J. Song and J. D. Hart, Bootstrapping in a high dimensional but very low-sample size problem, J. Stat. Comput. Simul. (to appear).

[26] J. D. Storey, A direct approach to false discovery rates, J. Roy. Statist. Soc. B 64 (2002), 479-498.

[27] J. D. Storey, The optimal discovery procedure: A new approach to simultaneous significance testing, UW Biostatistics Working Paper Series, Working Paper 259, 2005.

[28] J. D. Storey, J. E. Taylot and D. Siegmund, Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach, J. Roy. Statist. Soc. B 66 (2004), 187-205.

[29] L. Wasserman and K. Roeder, Weighted hypothesis testing, Technical Report, Carnegie Mellon University, 2006.

[30] B. L. Welch, The significance of the difference between two means when the population variances are unequal, Biometrika 29 (1938), 350-362.

[31] A. B. Wout, G. K. Lehrma, S. A. Mikheeva, G. C. O'Keeffe, M. G. Katze, R. E. Bumgarner, G. K. Geiss and J. I. Mullins, Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4+-*T*-cell lines, J. Virology 77 (2003), 1392-1402.

[32] M. Zheng, L. O. Barrera, B. Ren and Y. N. Wu, ChIP-chip: Data, model, and analysis, Biometrics 63(3) (2007), 787-796.