Communications for Statistical Applications and Methods
2018, Vol. 25, No. 5, 559–568

https://doi.org/10.29220/CSAM.2018.25.5.559
Print ISSN 2287-7843 / Online ISSN 2383-4757

Robust inference with order constraint in microarray study

Joonsung Kang[1,a]

[a]Department of Information Statistics, Gangneung-Wonju national University, Korea

Abstract

Gene classification can involve complex order-restricted inference. Examining gene expression pattern across groups with order-restriction makes standard statistical inference ineffective and thus, requires different methods. For this problem, Roy's union-intersection principle has some merit. The M-estimator adjusting for outlier arrays in a microarray study produces a robust test statistic with distribution-insensitive clustering of genes. The M-estimator in conjunction with a union-intersection principle provides a nonstandard robust procedure. By exact permutation distribution theory, a conditionally distribution-free test based on the proposed test statistic generates corresponding p-values in a small sample size setup. We apply a false discovery rate (FDR) as a multiple testing procedure to p-values in simulated data and real microarray data. FDR procedure for proposed test statistics controls the FDR at all levels of α and π_0 (the proportion of true null); however, the FDR procedure for test statistics based upon normal theory (ANOVA) fails to control FDR.

Keywords: classification, distribution-free test, false discovery rate, M-estimator, union-intersection principle

1. Introduction

Classification in a genome-wide study separates subjects into similar groups so that subjects in the same group are more similar to each other subjects in different groups (Kim and Park, 2015; Choi *et al.*, 2016). Gene expression data often has many outliers and is apt to be noisy. The small sample size in a microarray experiment makes the estimation of variance untrustworthy. A large number of genes and few number of arrays as well as a high signal-noise ratio in the microarray data make classical statistical approaches worthless. Robust statistical methods (Huber, 1981) offer a solution for the problem, especially when an underlying distribution is unknown. A robust statistical procedure should not be affected by departures from underlying assumptions caused by outliers (Jang *et al.*, 2018). That is, it performs well under underlying assumptions whereas its performance deteriorates as the situation gets different from the assumptions (Son and Kim, 2017). Particularly, among robust estimators, M-estimator performs well even when we have a small sample size (Lim, 2018).

Order-restricted inference has been an issue that has been investigated for the last sixty years (Robertson *et al.*, 1988). In a huge number of correlated heterogeneous genes with a small sample such as microarray data, order-restricted inference issues are often present in complicated ways. As for gene expression levels, the inference involves the mean expression over time or dose by inequalities of order-restriction (Peddada *et al.*, 2003). For example, as for k^{th} gene in the G groups, we can formulate hypotheses H_{0k} vs H_{1k} as below.

$$H_{0k} : \mu_{1k} = \mu_{2k} = \cdots = \mu_{Gk} \quad \text{vs} \quad H_{1k} : \mu_{1k} \leq \mu_{2k} \leq \cdots \leq \mu_{Gk},$$

[1] Department of Information Statistics, Gangneung-Wonju National University, Jukheon-gil 7, Gangneung-si 25457, Republic of Korea. E-mail: mkang@gwnu.ac.kr

Published 30 September 2018 / journal homepage: http://csam.or.kr
© 2018 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

where μ_{jk} means the average gene expression level in the k^{th} gene in the j^{th} group, where $j = 1, \ldots, G$.

Now we define union-intersection principle (Roy, 1953) to involve order-restricted inference. We have a null hypothesis $H_0 : \theta \in \Theta_0$ and rejection region C. A set of null hypotheses $\{H_l : \theta \in \Theta_l\}$ and resulting set of tests with rejection regions $\{C_l\}$, where l belongs to Ł, follows the union-intersection principle (UIP) if $\Theta_0 = \cap_{l\in Ł}$ and $C = \cup_{l\in Ł} C_l$ for each $l \in Ł$. Examining the gene expression pattern across different treatment groups requires different methods because the order limit makes standard statistical inference useless.

For this reason, nonstandard robust methods need to be addressed in order to classify genes taking order-restricted inference into account without assumptions. We propose an M-estimator based on a union-intersection principle for the distribution-insensitive classification of genes. An exact permutation enables a conditionally distribution-free test to compute p-values that are amenable in a small sample size, computationally tractable, and statistically robust.

Based on p-values, we need to test which genes has monotone increasing or decreasing pattern across the groups. Classical method in multiplicity controls the family-wise error rate (FWER), the probability of committing a type I error rate among all hypotheses at a preassigned level α. However, it is in general worthless to utilize it with a huge number of correlated gene due to losing truly differentially expressed genes or others. Benjamini and Hochberg (1995) proposed the false discovery rate (FDR) as an alternative of an the expected proportion of Type I error among the rejected hypotheses (genes). FDR procedures tend to produce less stringent Type I error compared to FWER controlling procedures (for example, the Bonferroni correction). Therefore, FDR procedures have greater power for the sake of increased numbers in Type I errors.

The paper is organized as follows. In Section 2, a general framework of robust inference is addressed. We discuss how to estimate $\mathbf{M}_n$ which is an M-estimator for a parameter of interest in each k^{th} gene. In Section 3, a union-intersection principle is used to construct the test statistics based upon $\mathbf{M}_n$ in Section 2. Then, we use the test statistics to compute p-values in conditionally distribution-free tests with a small sample size setup. In Section 4, simulated data and real microarray data are assessed by applying a multiple testing procedure (FDR). Section 5 provides the concluding remarks.

2. Robust inference

2.1. Data structure

There are n subjects across G groups in the K genes (positions). Each subject has a gene expression level. We take the linear model $\mathbf{Y}_k = \mathbf{X}\boldsymbol{\beta}_k + \mathbf{E}_k$ in the k^{th} gene into account where $\mathbf{Y}_k = (Y_{1k}, \ldots, Y_{nk})^t$ is the vector of gene expression levels collected from n individuals across G groups in the k^{th} gene and A^t denotes the transpose of a matrix A. The design matrix of the $n \times G$ matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 1 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

The regression parameter of the $G \times 1$ matrix is $\boldsymbol{\beta}_k = (\mu_{1k}, \delta_{2k}, \delta_{3k}, \ldots, \delta_{Gk})^t$ where μ_{1k} denotes the average gene expression level in the k^{th} gene in the first group and δ_{jk} denotes the difference between μ_{1k} and the average expression level in the k^{th} gene in the j^{th} group, $j = 1, 2, \ldots, G$. The vector of error of the $n \times 1$ matrix denotes $\mathbf{E}_k = (\epsilon_{1k}, \epsilon_{2k}, \ldots, \epsilon_{nk})^t$. We estimate a parameter of interest $\boldsymbol{\beta}_k$ for each $k = 1, \ldots, K$ with the following M-estimator.

2.2. M-estimator

Let $\rho : \mathfrak{R}_G \times X \to \mathfrak{R}$ be a measurable function. We define an M-estimator $\mathbf{M}_n$ as a solution by minimizing with respect to $\mathbf{t} \in \mathfrak{R}_G$.

$$\sum_{i=1}^{n} \rho\left(Y_i - \mathbf{x}_i^t \mathbf{t}\right),$$

where Y_i is the i^{th} observation of $\mathbf{Y}_k$ and $\mathbf{x}_i (= (x_{i1}, \ldots, x_{iG})^t)$, $i = 1, \ldots, n$ denotes the i^{th} row of $\mathbf{X}$. $\mathbf{M}_n$ is defined to be regression equivalent if $\mathbf{M}_n(\mathbf{Y} + \mathbf{Xb}) = \mathbf{M}_n(\mathbf{Y}) + \mathbf{b}$ for b in $\mathfrak{R}_G$. We define $\mathbf{M}_n$ to be scale equivalent if $\mathbf{M}_n(c\mathbf{Y}) = c\mathbf{M}_n(\mathbf{Y})$ for $c > 0$. Generally, the second condition is not satisfied. Fortunately, studentization makes $\mathbf{M}_n$ regression and scale equivalent. We consider a studentized M-estimator $\mathbf{M}_n$ of $\boldsymbol{\beta}_k$ as a solution for minimization

$$\sum_{i=1}^{n} \rho\left(\frac{Y_i - \mathbf{x}_i^t \mathbf{t}}{S_n}\right)$$

with respect to $\mathbf{t}$ ($G \times 1$ matrix) and $S_n = S_n(\mathbf{Y})$ is a scale statistic. The above linear model refers to the classical ANOVA model. But the distribution $Y_{1k}, \ldots, Y_{nk}$ may not be Gaussian. Researchers may be interested in G groups that may be stochastically ordered. For example, it is applicable to dose-response gene expression microarray data, introduced by Peddada *et al.* (2003) in cases when gene expression is stochastically increasing over time or dose. We define the null hypothesis H_{0k} as the fact that the G groups in the k^{th} gene are statistically homogeneous. The alternative hypothesis H_{1k} is because G groups in the k^{th} gene are ordered in an increasing level of dominance. H_{0k} and H_{1k} is constructed as below.

$$H_{0k} : \delta_{2k} = \delta_{3k} = \cdots = \delta_{Gk} = 0 \quad \text{vs} \quad H_{1k} : 0 \leq \delta_{2k} \leq \delta_{3k} \leq \cdots \leq \delta_{Gk}.$$

These can be restated as the following hypotheses.

$$H_{0k} : \boldsymbol{\theta}_k = \mathbf{A}\boldsymbol{\beta}_k = 0 \qquad H_{1k} : \boldsymbol{\theta}_k = \mathbf{A}\boldsymbol{\beta}_k \geq 0,$$

where the $(G-1) \times G$ matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & 0 & \cdots \\ 0 & 0 & -1 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$

So as to test the null hypothesis, we tend to take alternatives that the vector $\boldsymbol{\theta}_k$ belongs to the positive orthant space $\mathfrak{R}^{+(G-1)}$ into account. As for the univariate case, we have an optimal UMP test.

However, we do not have an optimal UMP test in a multivariate case. For instance, the Hotelling T^2 creates a huge set of confidence intervals and involves loss of efficiency. It is therefore interesting to consider statistical inference under restricted setup. We can utilize UIP (Roy, 1953) for such a statistical inference under the positive orthant multivariate alternative hypothesis. We take the first derivative of ρ function as ψ. $\mathbf{M}_n$ should be a median-unbiased estimator of $\boldsymbol{\beta}_k$. Symmetry of F and skew symmetry of ψ, which is defined to be the symmetry of top left with bottom right and top right with bottom left, are necessary conditions for the median-unbiasedness of $\mathbf{M}_n$. We select Huber loss function as a good candidate for ψ function. Therefore, minimization makes the estimator regression and scale equivalent. We define the Huber function as follows.

$$\rho(t) = \begin{cases} c|t| - \dfrac{1}{2} \times c^2, & \text{if } |t| > c, \\ \dfrac{1}{2} \times t^2, & \text{if } |t| \leq c. \end{cases}$$

The derivative of the Huber function ψ is defined as follows.

$$\psi(t) = \begin{cases} c \times \text{sign}(t), & \text{if } |t| > c, \\ t, & \text{if } |t| \leq c. \end{cases}$$

ψ function is decomposed into the sum

$$\psi = \psi_a + \psi_b + \psi_c.$$

ψ_a is absolutely continuous function with absolutely continuous derivative. ψ_c should be a continuous and piecewise linear function. ψ_b is an increasing step function. We have $\psi_a = \psi_b$ for the case of Huber loss function. This function satisfies the following conditions (Jurečková and Sen, 1996).

- M1: S_n should be both regression invariant and scale invariant, $S_n > 0$ a.s. and $n^{1/2}(S_n - S) = O_p(1)$.
- M2: $H(t) = \int \rho((z-t)/S)dF(z)$ has the singular minimum at $t = 0$.
- M3: For some $\delta > 0$ and $\eta > 1$,

$$\int_{-\infty}^{\infty} \left[|z| \sup_{|u| \leq \delta} \left| \psi_a'' \left(\frac{e^{-v}(z+u)}{S} \right) \right| \right]^{\eta} dF(z) < \infty$$

  and

$$\int_{-\infty}^{\infty} \left[|z|^2 \sup_{|u| \leq \delta} \left| \psi_a'' \left(\frac{e^{-v}(z+u)}{S} \right) \right| \right]^{\eta} dF(z) < \infty,$$

  where $\psi_a'(z) = (d/dz)\psi_a(z)$ and $\psi_a''(z) = (d^2/dz^2)\psi_a(z)$.
- M4: $\psi_c(z)$ should be a continuous and piecewise linear function with knots at $\mu_1, \ldots, \mu_r$. Henceforth the derivative $\psi_c'(z)$ is a step function

$$\psi_c'(z) = \alpha_v, \quad \mu_v < z < \mu_{v+1}, \ v = 0, 1, \ldots, r,$$

where $\alpha_0, \alpha_1, \ldots, \alpha_r \in \mathfrak{R}_1$, $\alpha_0 = \alpha_r = 0$, and $-\infty = \mu_0 < \mu_1 < \cdots < \mu_r < \mu_{r+1} < \infty$. $f(z)$ is assumed to be bounded in a neighborhood of $S_{\mu_1}, \ldots, S_{\mu_r}$.

- M5: $\psi_b(z) = \lambda_\nu$ for $q_\nu < z \le q_{\nu+1}, \nu = 1, \ldots, m$ where $-\infty = q_0 < q_1 < \cdots < q_{m+1} = \infty$, $-\infty < \lambda_0 < \lambda_1 < \cdots < \lambda_m < \infty$. $f(z)$ and $f'(z)$ are assumed to be bounded in $S_{q_1}, \ldots, S_{q_m}$. We represent $\mathbf{M}_n$ asymptotically in the following functionals

$$\gamma_1 = S^{-1} \int_{-\infty}^{\infty} \left(\psi_a'\left(\frac{z}{S}\right) + \psi_c'\left(\frac{z}{S}\right) \right) dF(z),$$
$$\gamma_2 = S^{-1} \int_{-\infty}^{\infty} z \left(\psi_a'\left(\frac{z}{S}\right) + \psi_c'\left(\frac{z}{S}\right) \right) dF(z).$$

Moreover, the following conditions are satisfied.

$$\text{X1. } x_{i1} = 1, \quad i = 1, \ldots, n,$$
$$\text{X2. } n^{-1} \sum_{i=1}^{n} \|\mathbf{x}_i\|^4 = O_p(1),$$
$$\text{X3. } \lim_{n\to\infty} \mathbf{Q}_n = \mathbf{Q},$$

where $\mathbf{Q}_n = n^{-1}\mathbf{X}^t\mathbf{X}$ and $\mathbf{Q}$ is a positive definite $p \times p$ matrix.

Under these conditions, $\mathbf{M}_n$ is a solution of the equation

$$\sum_{i=1}^{n} \mathbf{x}_i \psi\left(\frac{Y_i - \mathbf{x}_i^t \mathbf{t}}{S_n}\right) = \mathbf{0}.$$

We calculate S_n follows in order to make S_n scale and regression invariant. Regression scores as defined below are used.

For $\alpha \in (0, 1)$, $\hat{\mathbf{a}}_n(\alpha) = (\hat{a}_{n1}(\alpha), \ldots, \hat{a}_{nn}(\alpha))^t$ should be the unique solution to maximize

$$\sum_{i=1}^{n} Y_i \hat{a}_{ni}(\alpha)$$

with the constraint

$$\sum_{i=1}^{n} x_{ij} \hat{a}_{ni}(\alpha) = (1-\alpha) \sum_{i=1}^{n} x_{ij}, \quad j = 1, \ldots, G.$$

Hajék (1965) proposed scores

$$a_n^*(R_i, \alpha) = \begin{cases} 0, & \text{if } \dfrac{R_i}{n} < \alpha, \\ R_i - n\alpha, & \text{if } \dfrac{R_i - 1}{n} < \alpha < \dfrac{R_i}{n}, \\ 1, & \text{if } \alpha < \dfrac{R_i - 1}{n}. \end{cases}$$

An increasing and square integrable function $\phi : (0,1) \to \Re_1$, $\phi(\alpha) = -\phi(1-\alpha)$, $0 < \alpha < 1$ is chosen. For a number $\alpha_0 (0 < \alpha_0 < 12)$, ϕ is assumed to be standardized

$$\int_{\alpha_0}^{1-\alpha_0} \phi^2(\alpha) d\alpha = 1.$$

We define regression scores by ϕ as

$$\hat{b}_{ni} = -\int_{\alpha_0}^{1-\alpha_0} \phi(\alpha) d\hat{a}_{ni}(\alpha), \quad i = 1, \ldots, n.$$

That is,

$$\hat{b}_{ni} = \begin{cases} n \int_{\frac{(R_i-1)}{n}}^{\frac{R_i}{n}} \phi(\alpha) a_n^{*'}(R_i, \alpha) d\alpha, & \text{if } \alpha \le \dfrac{(R_i-1)}{n} \le 1-\alpha,\ \dfrac{R_i}{n} \le 1-\alpha, \\ n \int_{\frac{(R_i-1)}{n}}^{1-\alpha} \phi(\alpha) a_n^{*'}(R_i, \alpha) d\alpha, & \text{if } \alpha \le \dfrac{(R_i-1)}{n} \le 1-\alpha,\ \dfrac{R_i}{n} > 1-\alpha, \\ n \int_{\alpha}^{\frac{R_i}{n}} \phi(\alpha) a_n^{*'}(R_i, \alpha) d\alpha, & \text{if } \alpha > \dfrac{(R_i-1)}{n},\ \dfrac{R_i}{n} \le 1-\alpha, \\ 0, & \text{if } 1-\alpha < \dfrac{(R_i-1)}{n}, \\ n \int_{\alpha}^{1-\alpha} \phi(\alpha) a_n^{*'}(R_i, \alpha) d\alpha, & \text{else.} \end{cases}$$

We define S_n as

$$n^{-1} \sum_{i=1}^{n} Y_i \hat{b}_{ni} = n^{-1} \mathbf{X}^t \hat{\mathbf{b}}_n.$$

3. Union-intersection principle and multiple testing

3.1. Construction of test statistics for each k gene

Suppose that γ_1 is not equal to zero. For each k gene, the asymptotic distribution of $\mathbf{M}_n$ is as follows.

Theorem 1. *The sequence*

$$n^{\frac{1}{2}} \hat{\gamma}_1 (\mathbf{M}_n - \boldsymbol{\beta}_k) + \hat{\gamma}_2 \left(\frac{S_n}{S} - 1 \right) \mathbf{e}_1$$

follows the G-dimensional Gaussian distribution $N_G(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1})$ asymptotically, where $\sigma^2 = \int_{-\infty}^{\infty} \psi^2 (z/S) dF(z)$ (Jurečková and Sen, 1996).

The asymptotic variance of $n^{1/2}\hat{\gamma}_1(\mathbf{M}_n - \boldsymbol{\beta}_k) + \hat{\gamma}_2(S_n/S - 1)\mathbf{e}_1$ is $\kappa_k (= (\hat{\gamma}_1)^{-2} \hat{\sigma}^2 \mathbf{Q}^{-1})$. The same factor $(\hat{\gamma}_1)^{-2}\hat{\sigma}^2$ does not affect the parts of the following $\mathbf{Z}_k$ and $\mathbf{V}_k$. The last term $\hat{\gamma}_2(S_n/S - 1)\mathbf{e}_1$ is ignored while deriving a test statistic. Based upon the M-estimator $\mathbf{M}_n$ with theorem 1, we now tend to utilize the UIP in order to formulate a robust M-test as follows.

$$H_{0k} : \boldsymbol{\theta}_k = \mathbf{A}\boldsymbol{\beta}_k = 0 \qquad H_{1k} : \boldsymbol{\theta}_k = \mathbf{A}\boldsymbol{\beta}_k \ge 0,$$

where the $(G-1)\times G$ matrix

$$\mathbf{A}=\begin{pmatrix}0 & 0 & 0 & 0 & 0 & \cdots\\ 0 & 1 & 0 & 0 & 0 & \cdots\\ 0 & -1 & 1 & 0 & 0 & \cdots\\ 0 & 0 & -1 & 1 & 0 & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -1 & 1\end{pmatrix}.$$

$\mathbf{Z}_k = \mathbf{A}\mathbf{M}_n$ and $\mathbf{V}_k = \mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^t$, where $\mathbf{M}_n$ is an estimator of $\boldsymbol{\beta}_k$. For $\varsigma = \{1,\ldots,G-1\}$ and every $a : a \in \varsigma$, we set a' its complement and $|a|$ its cardinality. For each $a : a \in \varsigma$, we partition $\mathbf{Z}_k$ and $\mathbf{V}_k$ as follows.

$$\mathbf{Z}_k = \begin{pmatrix}\mathbf{Z}_{ka}\\ \mathbf{Z}_{ka'}\end{pmatrix},$$
$$\mathbf{V}_k = \begin{pmatrix}\mathbf{V}_{kaa} & \mathbf{V}_{kaa'}\\ \mathbf{V}_{ka'a} & \mathbf{V}_{ka'a'}\end{pmatrix}.$$

Write

$$\mathbf{Z}_{ka:a'} = \mathbf{Z}_{ka} - \mathbf{Z}_{ka'}^t\mathbf{V}_{ka'a'}^{-1}\mathbf{Z}_{ka'},$$
$$\mathbf{V}_{kaa:a'} = \mathbf{V}_{kaa} - \mathbf{V}_{kaa'}\mathbf{V}_{ka'a'}^{-1}\mathbf{V}_{ka'a}.$$

By virtue of weak convergence of $n^{1/2}(\mathbf{M}_n - \boldsymbol{\beta}_k)$ to a G-variable Gaussian law, for n very large, we derive

$$\left(n\mathbf{V}_k^{-1}\right)^{\frac{1}{2}}(\mathbf{Z}_k - \boldsymbol{\theta}_k) \xrightarrow{D} N_{G-1}(\mathbf{0},\mathbf{I}).$$

Then, we construct the proposed test statistic

$$L_k = \sum_{a\in\varsigma} I\left(\mathbf{Z}_{ka:a'} > \mathbf{0}, \mathbf{V}_{ka'a'}^{-1}\mathbf{Z}_{ka'} \le \mathbf{0}\right)\left(n\mathbf{Z}_{ka:a'}^t\mathbf{V}_{kaa:a'}^{-1}\mathbf{Z}_{ka:a'}\right).$$

3.2. p-value computation

We deal with high dimension low sample size data. n is small and we do not have asymptotic normality. The permutation distribution theory is still effective for such a setup. Under the null hypothesis of homogeneity, the joint distribution of n observations should be invariant under any permutation. We take all possible $n!/(n_1!n_2!\cdots n_G!)$ as equally likely permutations. Henceforth, we construct conditionally distribution-free tests by utilizing the permutation. We calculate p-value for each k gene as below.

$$P_k = \Pr(L_k \ge l_k), \quad k = 1,\ldots,K,$$

where L_k is a test statistic and l_k is an observed test statistic.

Table 1: Multiple hypothesis testing

	Not rejected	Rejected	Total
True null	U	V	m_0
Non-true null	T	S	$m - m_0$
	$m - R$	R	m

3.3. False discovery rate

We apply a proper multiple testing procedure (FDR) to K p-values to select which genes have monotone increasing or decreasing patterns.

FDR is defined as $E(V/R|R > 0) \cdot P(R > 0)$ in Table 1. V denotes the number of genes declared as differentially expressed among truly non-differentially expressed genes whereas S indicates the number of genes declared to be differentially expressed among truly differentially expressed genes. FDR has been in particular effective as the first alternative to the FWER to have broad acceptance in many scientific areas such as the life sciences, biology, chemistry, and medicine. For more details about the FDR, please see Dudoit *et al.* (2003), Storey (2002), and Sarkar (2006). For simulation study and real data analysis in next section, we select Benjamini and Hochberg's FDR procedure (Benjamini and Hochberg, 1995) and Storey's FDR procedure (Storey, 2002) among various FDR controlling procedures.

4. Numerical study

4.1. Simulation study

Simulation studies are conducted to show the performance of the proposed method and compare with the common procedure: ANOVA. We generate 100 random positions. Each position constitutes 5 groups with 30 observations. Each group contains 6 Gaussian variables with different means from others. Mean is assigned to each group in increasing order. The difference in the means between two adjunct groups increases by 1. We compute $\mathbf{M}_n$ based on those gene expression levels and corresponding L_k for each random position $k = 1, \ldots, 100$ by using a union-intersection principle in Section 3 that also enable computing the proposed test statistics. We then calculate 100 p-values via the proposed method by permuting the distribution of test statistics with about $30!/(6!)^5$ iterations. For comparison, we also compute 100 p-values via test statistics using ANOVA. We now determine which positions have monotone increasing patterns across the groups.

FDR is a method to compute the rate of type I errors in multiple testing. FDR-controlling procedure is used to control the expected proportion of discoveries (rejected hypotheses) that are false. As a multiple testing procedure, we apply Storey's FDR procedure and Benjamini and Hochberg's FDR procedure to a set of p-values. Storey denotes Storey's FDR procedure and BH denotes Benjamini and Hochberg's FDR procedure. Table 2 shows that for each α (significance level), Storey (proposed) and BH (proposed) are less than α for all π_0 (the proportion of true null hypotheses), which means they control the FDR at all levels of α and π_0. Storey (normal); in addition, BH (normal) using the existing test statistics (ANOVA) fail to control the FDR since they are more than α for some α and π_0 (Table 2). We infer that the proposed test statistics are optimistic for controlled FDR procedures and adequately reflect a monotone increasing or decreasing pattern of the response variable; however, existing test statistics fail to include a monotone increasing or decreasing trend of the response variable (failing to reflect genuine feature of the response variable).

Table 2: False discovery rate (1)

α	π_0	Storey (proposed)	BH (proposed)	Storey (normal)	BH (normal)
0.10	0.20	0.050	0.057	0.074	0.069
	0.40	0.063	0.077	0.089	0.098
	0.60	0.087	0.085	**0.105**	**0.102**
0.05	0.20	0.024	0.035	0.035	0.041
	0.40	0.036	0.049	**0.051**	**0.052**
	0.60	0.044	0.039	**0.055**	**0.054**
0.01	0.20	0.007	0.009	0.009	0.008
	0.40	0.008	0.008	**0.010**	**0.011**
	0.60	0.009	0.009	**0.011**	**0.012**

Storey = Storey's FDR procedure; BH = Benjamini and Hochberg's FDR procedure; FDR = false discovery rate.

Table 3: False discovery rate (2)

α	Storey (proposed)	BH (proposed)	Storey (normal)	BH (normal)
0.10	0.089	0.077	**0.110**	0.099
0.05	0.039	0.047	0.049	**0.053**
0.01	0.009	0.007	**0.013**	**0.015**

Storey = Storey's FDR procedure; BH = Benjamini and Hochberg's FDR procedure; FDR = false discovery rate.

4.2. Real data analysis

Application to the response of human fibroblasts to serum data set (Iyer *et al.*, 1999) introduced the chronological program at 12 time points of gene expression levels throughout the physiological response of fibroblasts to serum using cDNA microarrays that included 8,613 genes over 24 hours. They had 517 genes whose expression levels differed concerning the stimulation of serum. They measured gene expression levels at 0, 0.25, 0.15, 1, 2, 4, 6, 8, 12, 16, 20, 24 hours after the stimulation of serum. Our data constitute 1000 genes measured at 6 time points $1, 2, 4, 6$, and 8. One gene has 6 groups with 6 observations per group. We take log-transformation of gene expression levels. In order to compute the proposed test statistics in Section 3, we calculate $\mathbf{M}_n$ based on these gene expression levels and the resulting L_k for each gene $k = 1, \ldots, 1000$ using a union-intersection principle in Section 3. We then compute 1000 p-values via the proposed method by permuting about $36!/(6!)^6$ iterations. We also calculate 1000 p-values using test statistics by ANOVA. We then apply FDR procedures to the p-values. Estimated $\hat{\pi}_0$ = is 0.34 due to Storey's FDR, which means that the proportion of no differentially expressed genes is 0.34 among all genes. Storey's FDR procedure and Benjamini and Hochberg's FDR procedure are then applied to them. Table 3 shows that for each α (significance level), values of Storey (proposed) and BH (proposed) are less than α, which means that they control FDR at all levels of α. Storey (normal) is more than α for some α (= $0.1, 0.01$) (Table 3). BH (normal) is more than α for some α (= $0.05, 0.01$) (Table 3). Storey (normal) and BH (normal) using the existing test statistics (ANOVA) do not control the FDR. To be more specific, we conclude that the proposed test statistics are well designed (optimal) to control FDR procedures and adequately follow a monotone increasing or decreasing pattern of the response variable. However, ANOVA test statistics do not account for the monotone increasing or decreasing trend of the response variable.

5. Concluding remarks

Classification of genes in a microarray data is often involved in and order-restricted constraint. Outlier arrays often make standard statistical inference useless. For this reason, we propose an M-estimator using a union-intersection principle to test which gene has a monotone increasing or decreasing pattern

across the groups. FDR procedures based upon proposed test statistics control the FDR at all levels of α and π_0, but FDR procedures for the test statistics from normal theory do not control the FDR at all levels of α and π_0. Future research can be designed to develop test statistics based on a pure small sample theory. We extended the large sample theory results of $\mathbf{M}_n$ to small sample case. Microarray studies include many high dimension low sample size cases; therefore, we need to thoroughly explore test statistics with a small sample theory. In addition, we specifically designed a union-intersection principle for a monotone increasing or decreasing pattern of gene expression levels (the response variable) and an inequality-oriented inference. It is therefore worth developing union-intersection principle for a general case of constrained inference such as a shape constraint. In addition, more robust statistical methods can be developed for other measures such as $L-$statistics and a minimum distance method after examining the influence function for each measure.

References

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B*, **57**, 289–300.

Choi D, Choi H, and Park C (2016). Classification of ratings in online reviews, *Journal of the Korean Data and Information Science Society*, **27**, 845–854.

Dudoit S, Shaffer JP, and Boldrick JC (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science*, **18**, 71–103.

Hajék J (1965). Extension of the Kolmogorov-Smirnov test to regression alternatives. In *Proceedings of Bernoulli-Bayes-Laplace Seminar* (L. LeCam, ed.), 45–60.

Huber PJ (1981). *Robust Statistics*, John Wiley & Sons, New York.

Iyer VR, Eisen MB, Ross DT, *et al.* (1999). The transcriptional program in the response of human fibroblasts to serum, *Sciences*, **283**, 83–87.

Jang E, Choi S, and Kim D (2018). Robust Bayesian beta regression analysis, *Journal of the Korean Data and Information Science Society*, **29**, 27–36.

Jurečková J and Sen PK (1996). *Robust Statistical Procedures, Asymptotics and Interrelations*, Wiley, New York.

Kim G and Park C (2015). Analysis of English abstracts in journal of the Korean data and information science society using topic models and social network analysis, *Journal of the Korean Data and Information Science Society*, **26**, 151–159.

Lim Y (2018). M-estimation of the long-memory parameter by Laplace periodogram, *Journal of the Korean Data and Information Science Society*, **29**, 523–532.

Peddada SD, Lobenhofer EK, Li L, Afshari CA, Weinberg CR, and Umbach DM (2003). Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference, *Bioinformatics*, **19**, 834–841.

Robertson T, Wright FT, and Dykstra RL (1988). *Order Restricted Statistical Inference*, Wiley series in probability and Statistics, New York.

Roy SN (1953). On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics*, **24**, 220–238.

Sarkar SK (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures, *The Annals of Statistics*, **34**, 394–415.

Son N and Kim M (2017). A study on robust regression estimators in heteroscedastic error models, *Journal of the Korean Data and Information Science Society*, **28**, 1191–1204.

Storey JD (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society: Series B*, **64**, 479–498.

Received July 17, 2018; Revised August 29, 2018; Accepted August 30, 2018