MATEC Web of Conferences **44**, 01089 (2016)
DOI: 10.1051/matecconf/20164401089

A Generalized Order-restricted Inference Methodology for Selecting and Clustering Genes

Rui Yin Liu[a]
College of Mathematics and System Science, Shenyang Normal University, Shenyang, 110034, P.R. China

Abstract. There are many methods for selecting and clustering genes according to their time-course or dose–response profiles. These methods all necessitate the assumption of a constant variance through time or among dosages. This homoscedasticity assumption is, however, seldom satisfied in practice. In this paper, via the application of Shi's (1994,1998) algorithms and a modified bootstrap procedure, we proposed a generalized order-restricted inference methodology for the same task without the homoscedasticity restriction. Simulation results show that our procedure can control the false positive rate and have some good qualities.
Keywords: level probability; $\bar{E}^2$ test; bootstrap sampling; PAVA algorithm

1 INTRODUCTION

In microarray experiments and dose-response studies, experimental conditions usually have some inherent orderings. The performance of statistical inference can be improved significantly if these information can be appropriately utilized in the inferencial procedure . Order restricted statistical inference is an efficient tool to use these ordering information. Depending on the particular practical situation, one can use different order restricted methodologies. The problem under consideration in this present paper, namely clustering and selection of genes, is one example of such method.

Suppose there are T time points denoted by $1,2,...,T$, and at each time point there are n_t observations, for each of G genes. Let Y_{igt} denote the ith expression measurement taken on gene g at time point t, $\bar{Y}_{gt}$ denote the sample mean of gene g at time point t and $\bar{Y}_g = (\bar{Y}_{g1}, \bar{Y}_{g2}, ..., \bar{Y}_{gT})'$. The unknown true mean expression level of gene g is $(\mu_{g1}, ..., \mu_{gT})'$ which is restricted by some partial ordering, where $E(\bar{Y}_{gt}) = \mu_{gt}$. Inequalities between the components of $\mu_g = (\mu_{g1}, \mu_{g2}, ..., \mu_{gT})'$ define the true profile for gene g.

We adopt the following examples of inequality profiles according to Peddada et al. [1]. For notational convenience, the subscript g is dropped.

- Null profile: $C_0 = \{\mu \in R^T : \mu_1 = \mu_2 = \mathrm{L} = \mu_T\}$.
- Monotone increasing profile (simple order):

$$C_1 = \{\mu \in R^T : \mu_1 \le \mu_2 \le \mathrm{L} \le \mu_T\} \quad (1)$$

(with at least one strict inequality). One may similarly define a monotone decreasing profile by replacing $\le$ by $\ge$ in Equation (1).

- Up-down profile with maximum at i (umbrella order):

$$C_2 = \{\mu \in R^T : \mu_1 \le \mu_2 \le \mathrm{L} \le \mu_i \ge \mu_{i+1} \ge \mathrm{L} \ge \mu_T\} \quad (2)$$

(with at least one strict inequality among $\mu_1 \le \mu_2 \le \mathrm{L} \le \mu_i$ and one among $\mu_i \ge \mu_{i+1} \ge \mathrm{L} \ge \mu_T$. Genes satisfying this profile have mean expression values non-decreasing in time up to time point i and non-increasing thereafter. One may similarly define a down-up profile with minimum at i.

- Cyclical profile with minima at $1, j, T$ and maxima at i, k:

$$C_3 = \{\mu \in R^T : \mu_1 \le \mathrm{L} \le \mu_i \ge \mu_{i+1} \ge \mathrm{L} \ge \mu_j \le \mu_{j+1} \le \mathrm{L} \le \mu_k \ge \mu_{k+1} \ge \mathrm{L} \ge \mu_T\} \quad (3)$$

(with at least one strict inequality among each monotone sub-profile). Cyclical profiles may be important in long time-course experiments where the mean expression value could oscillate.

Our objective is to match the true profile of a gene, estimated from the observed data, to one of a specified set of candidate profiles. Peddada et al. [1] proposed a powerful method based on order-restricted inference, where the estimation procedure uses the known inequalities among parameters. They called the algorithm as ORIOGEN (Order Restricted Inference for Ordered Gene ExpressioN). But, their method necessitates the assumption of a constant variance through time. This homoscedasticity assumption is seldom satisfied in practice.

In fact, Simmons and Peddada [2] applied Hartley's test for heteroscedasticity of variances for the same microarray data of Lobehofer et al. [3]. As the Hartley's test is sensitive to normality assumption, and the gene expression data are not necessarily normally distributed, Simmons and Peddada [2] computed the P-values for the Hartley's test statistic by bootstrapping the residuals. A total of 367 (610) genes out of 1900 were heteroscedastic at a level of significance 0.05 (0.10).

Under the heterogeneous set up, Simmons and Peddada [2] provided a modified method, called ORIOG EN-Hetero. This algorithm allows heteroscedasticity by boot-strapping the residuals, that is, for each gene g and time point t obtain the residuals $e_{igt} = y_{igt} - \bar{y}_{gt}$, $i = 1,...,n_t$, $t = 1,...,T$. Next within the t^{th} time point draw a simple random sample of size n_t (with replacement) from $\{e_{1gt}, ..., e_{n_t gt}\}$. Denote the resampled residuals by $\{e^*_{1gt}, ..., e^*_{n_t gt}\}$. Then the bootstrap data y^*_{igt} are obtained by $y^*_{igt} = \bar{y}_g + e^*_{igt}$, where $\bar{y}_g = \sum_{t=1}^{T} n_t \bar{y}_{gt} \Big/ \sum_{t=1}^{T} n_t$. They compared its performance with the ORIOGEN whereas resampling was performed by mixing samples from all time points for a given gene, under the presence of heterogeneity, through a simulation study. They analyzed the microarray data of Lobehofer et

[a] Corresponding author: liury683@126.com

al. [3] using both ORIOGEN and ORIOGEN-Hetero, and observed that 197 out of 1900 genes were statistically significant at a level of significance $\alpha = 0.05$ using ORIOGEN, whereas 140 out of 1900 were significant by ORIOGEN-Hetero at the same level of significance, 115 being significant by both the methods.

In this present paper, we propose an alternative methodology to select and cluster genes, also by using the idea of order-restricted inference. For a given candidate profile, we estimate the mean expression level using the pool-adjacent-violators algorithm (PAVA), first published by Ayer et al. [4], see also Robertson et al. [5]. See the appendix for a brief description of this algorithm. The best fitting profile for a given gene is then selected using the goodness-of-fit criterion and the $\bar{E}^2$ test statistics (the likelihood ratio test statistics under unknown variance) under homoscedastic situation (see, Robertson, et al. [5]). This algorithm lead to a substantial gain of selection and clustering correction under monotone situation.

Also we propose another method based on the algorithm proposed by Shi [6] and Shi and Jiang [7] to deal with the heteroscedastic situation where a bootstrap technology is utilized to get samples under heteroscedastic situation.

The rest of the paper is organized as follows. In Section 2, we describe the proposed methodologies. Results from the simulation studies are presented in Section 3. In Section 4, an illustration of the proposed methodologies and a comparison with the existing methods are made using the real microarray experiment data of Lobenhofer et al. [3]. Section 5 concludes.

2 PROPOSED METHODOLOGY

2.1 Methodology under homoscedastic situation

For $t = 1,...,T$ and $i = 1,...,n_t$, suppose Y_{it} is a normally distributed random variable with unknown mean μ_t and variance of the form $\sigma_t^2 = a_t\sigma^2$, where $a_1,...,a_T$ be positive and known, and σ^2 be unknown. Also assume that Y_{it}'s are independent. The likelihood function is $(2\pi\sigma^2)^{-N/2} \cdot \prod_{t=1}^{T} a_t^{-n_t/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{T} a_t^{-1}\sum_{i=1}^{n_t}(y_{it}-\mu_t)^2\right]$ with $N = \sum_{t=1}^{T} n_t$.

It is well known that the vector $\bar{Y} = (\bar{Y}_1,...,\bar{Y}_T)'$ is the un-restricted maximum likelihood estimator (MLE) of $\mu = (\mu_1,...,\mu_T)'$, the population mean vector. Suppose, however, that it is known that $\mu \in C$, $C = \{C_j, j = 1,..., p\}$, a collection of candidate profiles, is defined as like in Section 1. Then one may want the estimate also satisfying that constraint, but due to sampling variability the sample means $\bar{Y}_1,...,\bar{Y}_T$ may not be ordered in this way. We obtain estimates satisfying this restriction by the isotonic regression Y^*. If Y^* minimizes $\sum_{t=1}^{T}[\bar{Y}_t - \mu_t]^2 \frac{n_t}{a_t}$ subject to $\mu \in C$, we call the restricted MLE Y^* the isotonic regression of $\bar{Y}$ with weights $w_t = \frac{n_t}{a_t}$. In the following we use $\hat{\mu}_C$ representing the isotonic regression of $\bar{Y}$ with respect to some order restriction defined in C. More specifically, $\hat{\mu}_{C_j}$ represents the isotonic regression of $\bar{Y}$ with respect to the order restriction C_j. For fixed $\sigma >$ 0, the isotonic regression $\hat{\mu}_C = (\hat{\mu}_{C_1},...,\hat{\mu}_{C_T})' \in C$ that maximizes $L(y;\mu,\sigma)$ is the same for each σ. For a fixed vector $\hat{\mu}_C$, the $\hat{\sigma}^2$ that maximizes $L(y;\mu,\sigma)$ is given by $\hat{\sigma}^2 = \sum_{t=1}^{T} a_t^{-1}\sum_{t=1}^{n_t}(y_{it} - \hat{\mu}_{C_t})^2 / N$.

Denoting the estimate of σ^2 given above (with $C = C_j$) by $\hat{\sigma}^2_{C_j}$, the likelihood ratio test (LRT) rejects C_0 against C_j for small values of $\Lambda = (\hat{\sigma}^2_{C_j} / \hat{\sigma}^2_{C_0})^{N/2}$. Equivalently, the LRT rejects C_0 for large values of $\bar{E}^2_{0j} = \frac{\sum_{t=1}^{T} w_t(\hat{\mu}_{C_j t} - \hat{\mu})^2}{\sum_{t=1}^{T} a_t^{-1}\sum_{i=1}^{M}(Y_{it} - \hat{\mu})^2}$, where $\hat{\mu} = \bar{y} = \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{n_t} y_{it}$ is the maximum likelihood estimate (MLE) of $\mu_t, t = 1,...,T$ under C_0, and $w_t = n_t / a_t$. Here $\hat{\mu}_{C_j} = (\hat{\mu}_{C_j 1},...,\hat{\mu}_{C_j T})'$ is the vector of the MLEs under C_j.

2.2 Algorithm I

First, let we rewrite some definitions given in Peddada et al. [1] in order to compare with their method.

Definition 1. Two parameters in a given profile are said to be linked if the inequality between them is specified *a priori*.

Definition 2. For a given profile, a parameter is said to be *nodal* if it is linked with every other parameter in the pro-file. For example, μ_i is the only nodal parameter in *profi-le* (2) while there are no nodal parameters in *profile* (3).

Definition 3. Define the l_∞ norm of an estimated profile as the maximum difference between the estimates of two linked parameters.

Definition 4. An inequality *sub-profile* C_i within a profile C is described by the inequalities between the compone-nts of the sub-vector $\mu_i = (\mu_{i_1},...,\mu_{i_s})$,where $\{i_1,...,i_s\} \subseteq \{1, ...,T\}$.

In view of the above definitions, Algorithm I consists of the following five steps:

Step 1. Pre-specify a collection of *p* candidate profiles. Denote these profiles by $C_1,...,C_p$.

Example 2.1 Suppose an experiment consists of four time points at 1, 2, 3, 4 hours, and we are interested in identif-ying genes belonging to either of the following profiles: $C_1 = \{\mu_1 \le \mu_2 \le \mu_3 \ge \mu_4\}$, $C_2 = \{\mu_1 \le \mu_2 \ge \mu_3 \ge \mu_4\}$.

For each gene $g, g = 1,...,G$, perform the following steps.

Step 2. Obtain the estimates of $\mu_{g1},...,\mu_{gT}$ under each of the candidate profiles $C_1,...,C_p$ using the PAVA algorithm (Barlow et al. [8] and Robertson et al. [5]).

Example 2.1 (continued). Suppose the sample mean expr-ession levels of a gene at the four time points are 1.8890, 2.8899, 1.8755, and 0.7342, respectively.

Estimation under C_1: Using the PAVA algorithm, we obtain $\hat{\mu}_1$ = 1.8890, $\hat{\mu}_2$ =2.3827, $\hat{\mu}_3$ = 2.3827, and $\hat{\mu}_4$ = 0.7324:

Estimation under C_2: In this case, $\hat{\mu}_1$ = 1.8890, $\hat{\mu}_2$ = 2.8899, $\hat{\mu}_3$ = 1.8755, and $\hat{\mu}_4$ = 0.7324.

Step 3. For each C_j, j = 1, 2,…, p, compute $\bar{e}_j^2$, where $\bar{e}_j^2 = \dfrac{\sum_{t=1}^{T} \frac{n_t}{a_t}(\hat{\mu}_{C_j t} - \bar{y})^2}{\sum_{t=1}^{T} a_t^{-1} \sum_{t=1}^{n_t}(y_{it} - \bar{y})^2}$. Compute $\max_j \bar{e}_j^2$.

Example 2.1 (continued). Here $\bar{e}_1^2$ = 0.3530, $\bar{e}_2^2$ = 0.4529, hence $\max(\bar{e}_1^2, \bar{e}_2^2)$ =0.4529.

Step 4 (BOOTSTRAP NULL DISTRIBUTION). Let the y_{it}'s be the observations at time t, t = 1,…, T. We draw N bootstrap samples. Each bootstrap sample is obtained as follows. Combine the y_{it} observations from all the time points into a vector of length $\sum_{t=1}^{T} n_t$ and draw T simple random samples y_{it}^* with replacement, each of size n_t. Repeat Steps 2 and 3 for each bootstrap sample y_{it}^*. This results in a bootstrap distribution for $\max_j \bar{e}_j^2$, which is used for testing $H_0 : \mu \in C_0, H_a : \mu \in \bigcup_{j=1}^{p} C_j$ (4)

Step 5. Assign gene g to profile C_r if $\bar{e}_r^2 = \max_j \bar{e}_j^2 \geq z_\alpha^*$, where z_α^* is the upper α th percentile of the bootstrap distribution derived in Step 4. If $\max_j \bar{e}_j^2 \leq z_\alpha^*$ or if two Profiles are tied then do not classify g into any of the p profiles.

Step 6. Repeat Steps 2-5 with every gene.

2.3 Methodology under heteroscedastic situation

The procedure described above is designed for genes with a constant variance over time or the situation where the unknown variances are in proportion. But it can not deal with situation where the unknown variances are subject to an order restriction or there are no conditions imposed on them (that is unrestricted and (possibly) unequal variances). Under this situation, we extend Peddada [1]'s method in two aspects. First, we obtain the estimates of the mean and variance using the algorithms proposed by Shi [6] and Shi and Jiang [7]. Second, we utilize a bootstrap technique to get samples under heteroscedastic situation.

2.4 Algorithm II

Step 1. Pre-specify a collection of candidate profiles. Denote these profiles by C_1, C_2,…, C_p.

For each gene g, g = 1, 2,…, G, perform the following steps.

Step 2. Obtain the estimates of $\mu_{g1},...,\mu_{gT}$ and $\sigma_{g1},...,\sigma_{gT}$ under each of the candidate profiles C_1,…, C_p using Shi's algorithms [5, 6].

Step 3. For each C_j , j = 1,…, p, compute $1_\infty^{g(j)}$. Let r be such that $1_\infty^{g(r)} = \max_j 1_\infty^{g(j)}$.

Step 4 (bootstrap null distribution). Let the y_{it}'s be the o-bservations at time t, t = 1,…, T, $\bar{y}_t$ and $s(y_t)^2$ the sample mean and the sample variance of y_{it}, respectively. We standardize y_{it} to $z_{it} = \dfrac{y_{it} - \bar{y}_t}{s(y_t)}$. Now the means and variances are same at every time point with respect to z_{it}. We draw N bootstrap samples. Each bootstrap sample is obtained as follows. Combine the z_{it} observations from all the time points into a vector of length $\sum_{t=1}^{T} n_t$ and draw T simple random samples z_{it}^* with replacement, each of size n_t. Then we transform z_{it}^* to y_{it}^* : $y_{it}^* = z_{it}^* s(y_t) + \bar{y}$,where $\bar{y}$ represents the sample mean of all observations. Repeat Steps 2 and 3 for each bootstrap sample y_{it}^*. This results in a bootstrap distribution for $\max_j 1_\infty^{r(j)}$, which is used for testing $H_0 : \mu \in C_0, H_a : \mu \in \bigcup_{j=1}^{p} C_j$ (4)

Step 5. Assign gene g to profile C_r if $1_\infty^{g(r)} \geq z_\alpha^*$, where z_α^* is the upper α th percentile of the bootstrap distribution derived in Step 4. If ` $1_\infty^{g(r)} \leq z_\alpha^*$ or if two profiles are tied then do not classify g into any of the p profiles.

Step 6. Repeat Steps 2-5 for every gene.

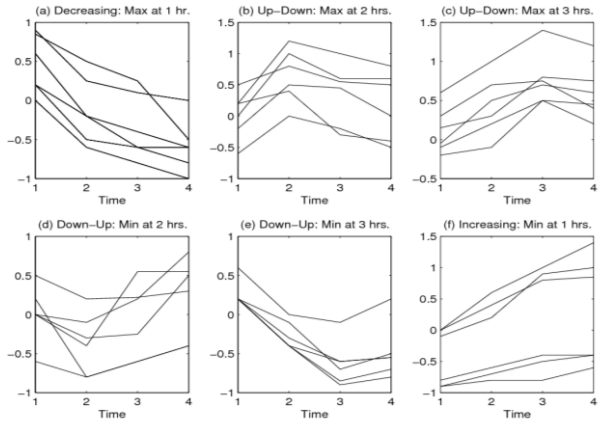

Figure1:Profiles of the simulated genes

3. SIMULATION RESULTS

We illustrate the proposed methodology, and compare the results with the results of Peddada et al. [1] and Simmons and Peddada [2], using gene expression data simulated from normal distributions. Suppose samples were harvested at 1, 2, 3, 4 hours. At each time point there are M = 15 observations. For each gene, we first assume that the variance of the expression was homoscedastic over time. For each gene g, we carry out the test procedure (4) where the alternative hypothesis is the union of the following six profiles: monotone decreasing C_1, monotone increasing C_6, two up-down profiles with maxima at 2, 3 hours C_2, C_3, respectively; and two down-up profiles with minima at 2, 3 hours C_4, C_5, respectively.

We generate two groups of data, from the same six types of profiles in Figure 1 with common variance 0.25 and 1 respectively. For every group we simulated *50* genes from every profile in Figure 1, in all *1800* genes. Assuming that the profiles of the genes are unknown, we wanted to select and cluster them to the six types of profiles using the above methodology.

Table 1(a). Selection and clustering results for data set 1.

l	ORIOGEN		Algorithm I		ORIOGEN-Hetero		Algorithm II	
	Selecting	Clustering	Selecting	Clustering	Selecting	Clustering	Selecting	Clustering
a	197	125	300	165	17	0	13	0
b	127	85	300	172	15	4	21	1
c	91	58	299	173	16	4	19	6
d	126	76	300	150	20	3	12	3
e	189	102	300	153	17	1	20	3
f	149	93	299	132	16	1	14	1

Using Algorithm I for the two groups of data with N = 10000, we selected 1798 (1800) genes with the test level α = 0.05, 300 (300) from genes simulated from Figure 1(a), 300 (300) from Figure 1(b), 299 (300) from Figure 1(c), 300 (300) from Figure 1(d), 300 (300) from Figure 1(e), 299 (300) from Figure 1(f), for the two variances. Of these, 165 (235) were clustered into C_1, 172 (238) into C_2, 173 (224) into C_3, 150 (220) into C_4, 153 (202) into C_5, 132 (180) into C_6.

We used the same data set with ORIOGEN algorithm with N = 10000 again. This time we selected 879 (1580) genes with the test level α =0.05, 197 (299) from genes simulated from Figure 1(a), 127 (248) from Figure 1(b), 91 (265) from Figure 1(c), 126 (233) from Figure 1(d), 189 (299) from Figure 1(e), 149 (236) from Figure 1(f). Of these, 125 (234) were clustered into C_1, 85 (195) into C_2, 58 (203) into C_3, 76 (182) into C_4, 102 (202) into C_5, 93 (156) into C_6.

Table 1(b). Selection and clustering results for data set 2.

l	ORIOGEN		Algorithm I		ORIOGEN-Hetero		Algorithm II	
	Selecting	Clustering	Selecting	Clustering	Selecting	Clustering	Selecting	Clustering
a	299	234	300	235	21	0	11	0
b	248	195	300	238	17	6	19	5
c	265	203	300	224	10	1	13	4
d	233	182	300	220	20	5	20	5
e	299	202	300	202	16	5	9	1
f	236	156	300	180	14	1	15	0

Using ORIOGEN-*Hetero* with N=10000, we selected 101 (98) genes with the test level α =0.05, 17 (21) from genes simulated from Figure 1(a), 15 (17) from Figure 1(b), 16 (10) from Figure 1(c), 20 (20) from Figure 1(d), 17 (16) from Figure 1(e), 16 (14) from Figure 1(f). Of these, 0 (0) were clustered into C_1, 4 (6) into C_2, 4 (1) into C_3, 3(5) into C_4, 1(5) into C_5, 1 (1) into C_6. See Tables 1(a) and 1(b) for these simulation results.

Using Algorithm II with N = 10000, we selected 99 (87) genes with the test level α =0.05, 13 (11) from genes simulated from Figure 1(a), 21 (19) from Figure 1(b), 19 (13) from Figure 1(c), 12 (20) from Figure 1(d), 20 (9) from Figure 1(e), 14 (15) from Figure 1(f). Of these, 0 (0) were clustered into C_1, 1 (5) into C_2, 6 (4) into C_3, 3 (5) into C_4, 3(1) into C_5, 1 (0) into C_6. See Tables 1(a) and 1(b) for these simulation results.

Clearly under all the situations from the results, ORIOGEN is inferior to our Algorithm I. Also we can see that Algorithm II and ORIOGEN-Hetero is inferior to both Algorithm I and ORIOGEN. The reason is obvious, because the simulated data is from homoscedastic variance, the weights ω in the algorithm obtaining the MLE estimate is fixed. But in Algorithm II and ORIOGEN-*Hetero*, the weight is computed with bootstrap samples in every iteration steps.

Table2. False positive rate of our procedure.

α	Case 1			Case 2			Case 3		Case 4
	Sc. 1	Sc. 2	Sc. 3	Sc. 1	Sc. 2	Sc. 3	Sc. 1	Sc. 2	Sc. 1
.06	.0547	.0572	.0512	.0531	.0559	.0510	.0668	.0557	.0610
.05	.0469	.0411	.0439	.0466	.0497	.0439	.0565	.0513	.0523
.03	.0227	.0247	.0216	.0208	.0249	.0219	.0326	.0285	.0392
.02	.0156	.0163	.0157	.0145	.0166	.0133	.0204	.0187	.0229
.01	.0087	.0090	.0075	.0073	.0081	.0065	.0118	.0094	.0131
.009	.0069	.0070	.0064	.0068	.0075	.0066	.0103	.0085	.0185
.006	.0046	.0044	.0047	.0041	.0049	.0044	.0076	.0056	.0056
.003	.0020	.0022	.0024	.0024	.0025	.0020	.0033	.0029	.0031

Also we generated 2000 genes from a profile with the common mean 0 across all the time points for both groups. We note that using Algorithm I and ORIOGEN with the test level α =0.05, we selected 94 and 104 genes, respectively, from the 1840 genes of the first group, 102 and 114 genes of the other group. Obviously, ORIOGEN resulted in more false positives.

We also investigated the false positive rate of our procedures. To generate unpatterned null data, we created 60 observations from four standard normal populations, each with 15 observations. Our simulations suggest that our methodology provides fairly accurate type I error rates.

Table 2 reports the false positive rates. We explored four cases. For both the Case 1 and Case 2, the null hypothesis is, H_0: $\mu_1 = \mathrm{L} = \mu_4$, the alternative H_c is taken as the following scenarios respectively. Scenario 1: $\mu_1 \le \mu_2 \le \mu_3 \ge \mu_4$ or $\mu_1 \le \mu_2 \ge \mu_3 \ge \mu_4$;Scenario 2: $\mu_1 \le \mu_2 \le \mu_3 \le \mu_4$ or $\mu_1 \le \mu_2 \le \mu_3 \ge \mu_4$; Scenario 3: $\mu_1 \le \mu_2 \le \mu_3 \le \mu_4$ or $\mu_1 \le \mu_2 \le \mu_3 \ge \mu_4$ or $\mu_1 \le \mu_2 \ge \mu_3 \ge \mu_4$.

We assume the variances are equal in Case 1, that is $\sigma_1^2 = \mathrm{L} = \sigma_4^2 \mathrm{A}\, \sigma^2$, and the data in simulation are from $N_4(0, I_4)$. In Case 2 the variances are in proportion $(\sigma_1^2, ..., \sigma_4^2)$

$=(a_1,...,a_2)\sigma^2$, the data in simulation are from $N(0, \text{diag}(1, 3.24, 4, 2.25))$. In Case 3, the variances are subject to some order restriction. In Scenario 1 the null hypothesis $H_0:\mu_1=\text{L}=\mu_4,\sigma_1=\text{L}=\sigma_4$, the alternative $H_a:\mu_1\le\mu_2\le\mu_3\ge\mu_4$, $\sigma_1\ge\sigma_2\ge\sigma_3\le\sigma_4$ or $\mu_1\le\mu_2\ge\mu_3\ge\mu_4,\sigma_1\ge\sigma_2\le\sigma_3\le\sigma_4$. In Scenario 2 we assume the variances are subject to a simple order $\sigma_1\ge\sigma_2\ge\sigma_3\le\sigma_4$ whatever be the null or alternative hypo theses. For this case, we discussed H_0: $\mu_1=\mu_2=\mu_3=\mu_4$ vs. H_a: $\mu_1\le\mu_2\ge\mu_3\ge\mu_4$ or $\mu_1\le\mu_2\le\mu_3\ge\mu_4$. In Case 4, we have no idea of the variances. Here the null hypothesis H_0 is $\mu_1=\mu_2=\mu_3=\mu_4$, the alternative $H_a:\mu_1\le\mu_2\ge\mu_3\ge\mu_4$ or $\mu_1\le\mu_2\le\mu_3\ge\mu_4$.

Table 3. False positive rate of Peddada's method.

α	.06	.05	.03	.02	.01	.009	.006	.003
Sc. 1	.060 8	.052 1	.026 1	.020 3	.010 6	.008 2	.005 0	.003 4
Sc. 2	.062 4	.053 6	.031 4	.017 7	.009 7	.008 4	.007 7	.002 0
Sc. 3	.060 1	.053 0	.032 6	.018 9	.009 3	.008 8	.005 2	.003 0

In contrast, we also investigated the false positive rate of the ORIOGEN method using bootstrap sampling under a constant variance through time. Only a small portion of the results are presented in Table 3, for the sake of brevity. Here the null hypothesis for all the cases is H_0: $\mu_1=\mu_2=\mu_3=\mu_4$, the alternative H_a is also taken as Scenarios 1, 2, and 3, respectively. It is necessary to point out that ORIOGEN based on bootstrap sampling is very sensitive to the observations themselves and the number of observations.

Table 4. False positive rates of Algorithm II and the ORIOGEN-Hetero method with 6 time points.

	Pattern (i=1,…,6)		Algorithm II	ORIOGEN-*Hetero*
1	$\mu_i=0$	$\sigma_i^2=16$	0.033	0.034
2	$\mu_i=0$	$\sigma_i^2=i^2$	0.037	0.044
3	$\mu_i=0$	$\sigma_i^2=i^3$	0.040	0.032
4	$\mu_i=i$	$\sigma_i^2=16$	0.289	0.330
5	$\mu_i=i$	$\sigma_i^2=i^2$	0.717	0.740
6	$\mu_i=i$	$\sigma_i^2=i^3$	0.291	0.307
7	$\mu_i=0,i=1,2,4,5,6,\mu_3=3$	$\sigma_i^2=16,i=1,2,4,5,6,\sigma_3^2=9$	0.605	0.602

We also carry out a simulation study based on 10,000 bootstrap samples and 1000 genes, at a significance level of 0.05. We compare the performances of false positive rates of the proposed method under heteroscedastic model in our paper (Algorithm II) and the ORIOGEN-Hetero algorithm of Simmons and Peddada [2]. The results are given in Table 4.

4. ANALYSIS OF REAL DATA USING DIFFERENT MRTHODOLOGIES

We consider the microarray experiment data set of Lobenhofer et al. [2] for illustration and comparison of different methodologies.

First we use Algorithm I and ORIOGEN to deal with the data assuming the homoscedasticity. At the significance level 0.005, we selected 223 out of 1900 genes using Alg-orithm I. From the selected genes, we obtained 15 genes belonging to the profile 1, 51 genes belonging to profile 2. Compared to Peddada et al. [1], there they selected 197 out of 1900 genes using ORIOGEN, there are 161 common genes. Because Algorithm I is sensitive to monotonic situation, and both 1 and 2 profile are monotonic, so Algorithm I selected 26 more genes than ORIGEN. This result tallies with the simulation.

Under heteroscedastic situation, we selected 132 genes using Algorithm II. Compared to ORIOGEN-Hetero, there are 108 genes in common. In ORIOGEN-Hetero the bootstrap sampling is restricted to one time point' residual, that is, we resample from only 8 observations at every time point t respectively. But in Algorithm II, we resample from all the observations at all the time points, that is 48 observations. There are 123 common genes between Algorithm I and ORIOGEN-Hetero. Between ORIOGEN and Algorithm II, there are 108 genes in common.

5. DISCUSSION

In this article, our research is mainly focused on two aspects of the selection and clustering algorithm. The first relates to the test method. Here we use an accurate test (i.e., LRT) to judge whether a gene should be rejected or accepted.

The second emphasis in our research is to explore the bootstrap sampling under heteroscedastic situation. We obtain samples from the standardized observations, and then take a transform to revert the samples so that they are from the null hypothesis. The simulation results show that this technique works well, and even works better than the ORIOGEN-Hetero algorithm of Simmons and Peddada [2].

Because this article focuses on the method for selection and clustering, we have not concentrated much on the analysis of the data. It should be pointed out, however, that our methodology enjoys several desirable properties. For example, the estimated mean expression levels in all cases, subject to an inequality profile, are MLEs.

In studies where experimental conditions have an inherent ordering, making use of ordering information can improve inference. In microarray experiments, the ability to exploit ordering information may be especially valuable as the genes whose expression levels change in concert through time may be components of the same cellular process or may share regulatory elements. Peddada et al. [1] have recognized the importance of time-course information and developed procedures based

on the statistical theory of order-restricted inference that makes explicit use of ordering information when selecting differentially expressed genes. Here we proposed a new selection and clustering method based on LRT under homoscedastic situation, and generalized Peddada's method to heteroscedastic situation. It is expected that methods of analysis that exploit the ordering of treatments to improve estimation will become increasingly valuable for time-course and dose-response micro-array experiments.

ACKNOWLEDGEMENTS

This work is supported by NSFC (Grant Numbers 11401393) and Liaoning Province Doctoral Startup Fund (Grant Number 20131107).

REFERENCES

1. Peddada S, Lobenhofer E, Li L, Afshari C, Weinberg C, Umbach D. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics, **19**: 834-841. (2003)
2. Simmons S, Peddada S, Order-restricted inference for ordered gene expression (ORIOGEN) data under heteroscedastic variances. Bioinformatics, **1(10)**: 414-419. (2007)
3. Lobenhofer E, Bennett L, Cable P, Li L, Afshari C. Regulation of DNA replication for k-genes by 17- β estradiol. *Molec*. Endocrin, **16**: 1215-1229. (2002)
4. Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E. An empirical distribution function for sampling with incomplete information. Ann. Math. Statist.,**26**: 641-647. (1955)
5. Robertson T, Wright FT, Dykstra RL. Order Restricted Statistical Inference. Wiley: New York, (1988).
6. Shi NZ. Maximum likelihood estimation of means and variances from normal populations under simultaneous order restrictions. J. Mult. Anal.; **50**: 282-293. (1994)
7. Shi NZ, Jiang H. Maximum likelihood estimation of isotonic normal means with unknown variances. J. Mult. Anal.; **64**: 183-195. (1998)
8. Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD. Statistical Inference under Order Restrictions. Wiley: New York, (1972).