2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Detection of non-structural outliers for microarray experiments

Zihua Yang and Zheng Rong Yang

***Abstract*— Outliers are unavoidable in many experiments due to various complex reasons ranging from equipment resolution to data contamination. The presence of outliers in microarray gene expression data can affect the quality of gene selection and ranking. This effect is severe when a microarray gene expression data is composed of too few samples. We classify outliers occurred in microarray gene expression data as structural and non-structural outliers. Structural outliers are gene dependent or sample dependent (or both) whereas non-structural outliers are gene and sample-independent. They are uninformative to gene expression differentiation but can cause misclassification of a differentially expressed gene as a non-differentially expressed one. While there are algorithms for detecting structural outliers, a different strategy is required for detecting non-structural outliers. We show the impact of non-structural outliers on gene selection/ranking and false discovery rate control. We also show the unsuitableness of existing outlier detection algorithms for detecting non-structural outliers. We propose a new algorithm for detecting non-structural outliers. It models the consecutive differences of ordered gene expressions as exponentially distributed. We use simulated and real data to demonstrate the efficacy of the proposed algorithm in correcting for non-structural outliers and improving gene selection/ranking and false discovery rate control.**

I. Introduction

LIKE many other experimental data, microarray gene expression data also contain outliers because of various complex reasons. The occurrence of outlier genes (genes whose expressions containing outliers) affects the overall accuracy of microarray gene expression data analysis such as gene selection/ranking [1]. Although the outlier problem in microarray data has been widely studied, the status of outliers has not been properly classified. We refer to outliers consistently found in one or a few arrays (samples) as sample-dependent structural outliers. They are often deemed erroneous and the removal of a complete array with sample-dependent structural outliers can therefore improve biomarker predictive capability [2-7]. Gene-dependent structural outliers are informative for gene expression differentiation and often studied in the context of heterogeneous differentially expressed genes [8-11]. Existing algorithms for detecting gene-dependent structural outliers (referred to as gene-specific outlier detection algorithms) include cancer profile outlier analysis (COPA) [12], outlier sum (OS) [13], outlier robust *t* statistic (ORT) [14] and maximum ordered subset *t* statistic (MOST) [15]. These algorithms account for the effect of gene-dependent structural outliers on the estimated pooled variance by adjusting the denominator of the *t* statistic for each gene. As outlier genes are predicted gene-by-gene by these gene-specific outlier detection (GOD) algorithms, sufficient number of samples (replicates) is required for robust inference and decision-making.

Zheng Rong Yang are with School of Biosciences, University of Exeter, UK (e-mail: z.r.yang@ex.ac.uk).

Zihua Ynag is with University of Queen Mary, UK (e-mail: z.h.yang@qmul.ac.uk)..

Non-structural outliers (NSOs) are distributed randomly across gene expressions. They do not inform gene expression differentiation but their presence can affect the identification of differentially expressed genes (DEGs) across conditions, e.g. from control to test conditions. For instance, the occurrence of a very low expression NSO to a test sample may make an originally up-regulated DEG misclassified as a non-DEG. The occurrence of a very high expression NSO to a control sample may also make an original up-regulated DEG misclassified as a non-DEG as well. NSOs thus have the impact on gene selection/ranking. In addition, the false discovery rate control relies on the quality of *p* values acquired from a gene selection/ranking process using such as *t* test or modified *t* test. If the distribution of *p* values obtained from a gene selection/ranking process is skewed from an expectation, false discovery rate control becomes difficult.

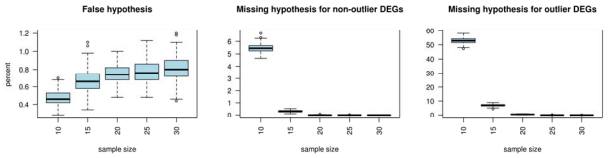

Fig. 1. FH and MH rates (%) for the study of how NSOs affect missing prediction of DEGs.

The existence of NSOs causes misclassified DEGs but not non-DEGs. We generated a data set with 5,000 up-regulated DEGs and 5,000 non-DEGs, with 10% outlier genes. Note that simulation using down-regulated DEGs will end up with the same conclusion because of the symmetrical property. The control samples and the test samples of non-DEGs were drawn from $\mathcal{N}(10,1)$. The test samples of DEGs were drawn from $\mathcal{N}(12,1)$. The number of samples was set to ten, 15, 20, 25 and 30. Low-expression outliers in test samples or high-expression outliers in control samples were randomly subtracted/added across the samples with outlier distances randomly drawn from $\mathcal{N}(2,0.1)$. We repeated the simulation 100 times. We estimated the probability that a DEG with an outlier was predicted as a non-DEG (missing hypothesis), the probability that a DEG without an outlier was predicted as a non-DEG (missing hypothesis) and the probability that a non-DEG was predicted as a DEG (false hypothesis). The outcome shows that both false hypothesis (FH) and missing hypothesis (MH) rates for non-outlier DEGs were low whereas the missing hypothesis rate for outlier DEGs was high (especially for small sample sizes) - Fig. 1.

It is also expected that large outlier percentage or small sample number will weaken gene selection/ranking accuracy. We therefore carried out a simulation with the same data used above but varying sample number from two to ten and

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

varying outlier gene percentage from 10% to 50%. The simulation proved this relationship - Fig. 2.

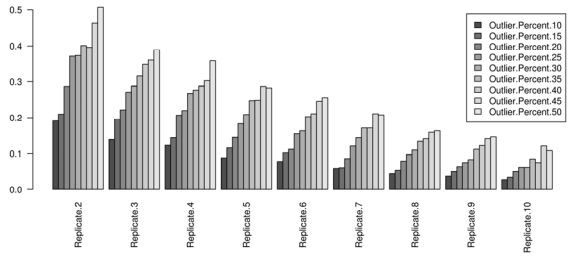

Fig. 2. The prediction accuracy drop from data without outlier genes to data with outlier genes inserted

The existence of NSOs leads to overestimation of null gene proportion and increased false discovery rate. Gene significance analysis (or gene selection/ranking) is typically carried out on multiple genes simultaneously (multiple hypothesis testing). In the absence of outliers, the resulting *p* values typically follow a mixture density, $g(p) = \pi_0 g_0(p) + (1-\pi_0) g_1(p)$, where π_0 is the null gene (non-DEGs) proportion, $g_0(p)$ is the uniform density of null genes and $g_1(p)$ is some fast decaying density of alternative genes (DEGs). A well-distributed mixture is the basis for proper false discovery rate control [16-19]. However, the existence of outlier genes distorts the *p* value distribution and in particular leads to overestimated null gene proportion π_0 and an under-estimated alternative gene proportion $1-\pi_0$. For small sample size microarrays, a few outliers can be sufficient to change a DEG to a non-DEG. To illustrate this, we simulated a data set with $\pi_0 = 0.5$ for 5,000 DEGs and 5,000 non-DEGs with and without outliers. All control samples and test samples of non-DEGs were drawn from $\mathcal{N}(10,1)$ and the test samples of DEGs were drawn from $\mathcal{N}(11,1)$. Outliers were added either by subtracting the minimum expression by a value drawn from $\mathcal{N}(2,0.1)$ or by adding a value drawn from $\mathcal{N}(2,0.1)$ to the maximum expression (random alternation between high and low expression outliers). We used eBayes [16] - a modified *t* test - to calculate *p* values. The null gene proportion π_0 is estimated using the method described in [18] $\hat{\pi}_0(\lambda) = \#\{p_i > \lambda\}/(1-\lambda)m$, where p_i is the *p* value calculated using eBayes for the i^{th} gene, *m* is the number of genes in a microarray gene expression data set, λ a pre-determined critical *p* value and $\hat{\pi}_0(\lambda)$ means the estimation of π_0 in terms of λ. A positive bias in the number of genes predicted as non-DEGs $\#\{p_i > \lambda\}$ then leads to an overestimated π_0 because the denominator of the above equation is fixed when λ and *m* are fixed. The simulation was repeated for 100 times for sample sizes of five and ten. It is evident from Fig. 3 that the π_0 increases along with the increase of outlier percentage. According to Storey [18], the positive false discovery rate (pFDR) is estimated using

$$\text{pFDR} = \frac{\hat{\pi}_0 \gamma}{\{\#\{p_i \le \gamma\} \vee 1\}\{1-(1-\gamma)^m\}}$$

where $[1, \gamma]$ is the rejection region and $\vee$ the OR operator (to avoid singularity). It can be seen that when π_0 is increased and $\#\{p_i \le \gamma\}$ is decreased, pFDR is increased if γ is fixed. This means that when the null gene population is overestimated, poorer false discovery rate control will happen.

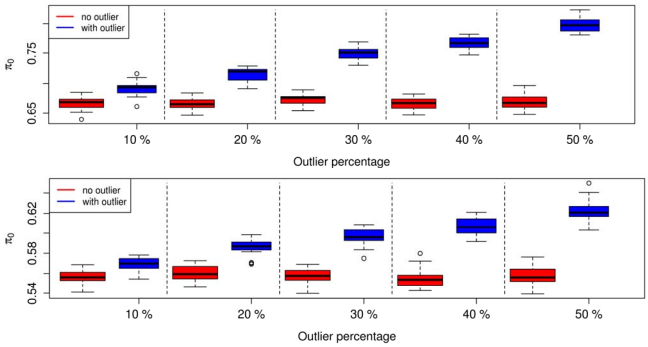

Fig. 3. Estimated π_0 for simulations with five samples (top) and ten samples (bottom) for varying outlier percentages. $\lambda = 0.01$.

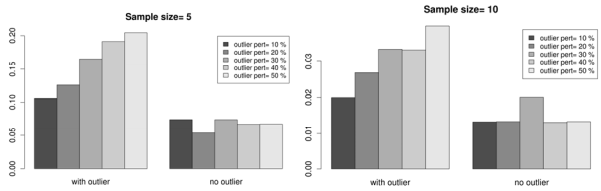

Fig. 4. Estimated Storey's mean pFDR for data (ten samples) with and without outliers for varying outlier percentages and sample sizes five and ten.

Fig. 4 shows the mean pFDR across all combinations of $\lambda \in \{0.001, 0.002, \cdots, 0.01\}$ and $\gamma \in \{0.01, 0.02, \cdots, 0.2\}$. It can be seen that data with outliers typically resulted in increasing pFDRs due to overestimated π_0. The difference was less noticeable when the sample size increased to ten.

In this paper, we introduce a new algorithm for detecting NSOs by modelling the ordered consecutive expression distances as an exponentially distributed at the population level (using whole microarray gene expression data). This exponential distance model assesses potential outliers in terms of distance to the non-outlier samples. In practice, we expect NSOs to constitute a minority of the population of pooled expressions, contributing a small amount to the tail of the overall (exponential-like) distribution of expressions. We name this new algorithm as POD standing for Population-based Outlier Detection. We illustrate the efficacy of this algorithm using simulated data and cancer microarray expression profile data. We also discuss how the power of gene selection/ranking and false discovery rate control can be enhanced through correcting detected outliers using a simple imputation approach.

II. ALGORITHM AND EXPERIMENTAL DESIGN

A. Algorithm

We denote a matrix of (log2) expressions by **X**, which has n rows for n genes and m columns for m samples. The expressions in **X** are sorted in an ascending order row by row, i.e. $x_{i,(1)} < x_{i,(2)} < \cdots < x_{i,(m)}$, $\forall i \in [1,n]$. Outlier detection is carried out separately for control and test samples. We then define a non-negative distance vector $\mathbf{z} = \{\delta_{ij}\}$, where

$$\delta_{ij} = x_{i,(j+1)} - x_{i,(j)}$$

$i \in [1,n]$ and $j \in [1,m-1]$, $x_{i,(j)}$ is the j^{th} smallest expression in the i^{th} row of **X**. We found the exponential distribution to fit δ_{ij} well in most cancer microarray data. The outlier p values can be evaluated using the fitted distance distribution. A gradient descent learning algorithm is used to find the optimal exponential rate parameter β with an initial value of $\beta = 1$,

$$\Delta\beta = -\eta\, \mathbf{e}\, \mathbf{G}\, (1 - \beta\, \mathbf{z}_0)$$

where $\eta = 0.1$ is an update coefficient and

$$\mathbf{e} = \mathbf{f}_0 - \beta\, e^{-\beta \mathbf{z}_0}$$

$\mathbf{f}_0$ is the empirical density (histogram) of $\mathbf{z}_0$ and $\mathbf{z}_0$ is the set of the middle points across the bins of $\mathbf{f}_0$. **G** is defined as $\mathrm{diag}(e^{-\beta \mathbf{z}_0})$. A non-negative distance is predicted as an outlier distance if the outlier p value is less than a pre-defined threshold, we set this to be 0.05. In the situation where more than one significant outlier distance arise for a gene, the distance with the minimum outlier p value is used to determine the boundary of the two subpopulations $\{x_{i,(1)}, x_{i,(2)}, \cdots, x_{i,(k)}\}$ and $\{x_{i,(k+1)}, x_{i,(k+2)}, \cdots, x_{i,(m)}\}$. We set an upper bound of 25% on the outlier proportion.

B. Simulated data

Two sets of simulated data were generated for the assessment of the proposed algorithm. The first set was composed of 10% of outlier genes and the other set was composed of 30% outlier genes. Both data sets were composed of 900 non-DEGs and 100 up-regulated DEGs. The samples of non-DEGs and the control samples of DEGs were drawn from $\mathcal{N}(10,1)$. The test samples of DEGs were drawn from $\mathcal{N}(11,1)$. Outlier genes were randomly selected. Each outlier gene was composed of one outlier. The distance (referred to as outlier distance) between non-outlier samples and an outlier sample of an outlier gene was drawn from $\mathcal{N}(\delta,0.1)$, where δ was one, two and three. For a selected gene, an outlier was randomly inserted into control or test samples. If an outlier was inserted into control samples, a high-expression outlier was used. If an outlier was inserted into test samples, a low-expression outlier was used. The sample number was five and ten. We designed two sets of accuracy measurements for DEG prediction and outlier gene prediction. 1) We used pFDR and sensitivity for examining how POD improves the prediction accuracy of DEGs. This means that we compared DEG prediction accuracy after outlier detection/correction against DEG prediction accuracy before outlier gene detection/correction. The meaning of pFDR has been aforementioned. Sensitivity was defined as the ratio of predicted DEGs over the designed DEGs. 2) We used AUR and sensitivity to evaluating how accurate POD and GOD algorithms identify outlier genes. AUR stands for area under ROC curve. ROC stands for receiver operating characteristic [20, 21] and is typically used as a robustness measure in two-class classification analysis tasks. A ROC curve describes how the sensitivity (also called the true positive rate) varies along with the false positive rate (also called the false alarm rate). Varying the cutting point for classification between non-outlier genes and outlier genes gives the multiple pairs of false positive rates and sensitivities on a ROC curve. A robust classifier is characterised by a ROC curve which is close to the top-left corner, or equivalently a large AUR. We used five GOD algorithms COPA [12], OS [13], ORT [14], MOST [15] and LSOSS [22] for comparison.

Table 1. Seven data sets downloaded from GEO and IGC.

Accession	Cancer type	No of probes	No of samples
gds1439	Prostate cancer	54675	7/6
gse12630	Breast cancer	22283	4/7
gse12630	Liver cancer	22283	4/4
gse7410	Cervical cancer	43931	16/13
IGC	Breast cancer	54675	53/4
IGC	Colon cancer	54675	26/20
IGC	Sarcoma cancer	54675	4/15

C. Cancer data

We downloaded seven data sets from GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) and IGC (International Genomics Consortium, http://www.intgen.org/) - Table 1.

D. Outlier correction

We used a simple imputation approach for outlier correction, i.e. using the mean of non-outliers to replace an identified outlier for an outlier gene. The significance analysis was carried out using eBayes (in the R package limma) and q value (R qvalue package). The late was used for false discovery rate control.

III. RESULTS

It was assumed that the detection and correction of outliers will turn over the misclassification of a DEG as a non-DEG. Therefore it was expected that the sensitivity of predicting DEG after outlier detection/correction should be improved compared with that before outlier detection/correction. From Fig. 5, we can find the following facts. *First*, the outlier detection/correction improved DEG prediction accuracy all the way. The improvement was obvious though it was still imperfect. *Second*, when the replicate number was larger, the prediction accuracy of DEGs was higher. *Third*, when the outlier distance was larger, the improvement of the DEG prediction accuracy was greater.

As analyzed above, outlier genes make contribution to increased pFDR. It was then expected that the correction of detected outliers should help reduce pFDR. It was also

expected that the impact of outlier genes on pFDR should be smaller when the sample size was larger. Fig. 6 illustrates these two facts very well, where we can see the consistent drop of pFDR between two stages, i.e. before and after outlier gene detection/correction.

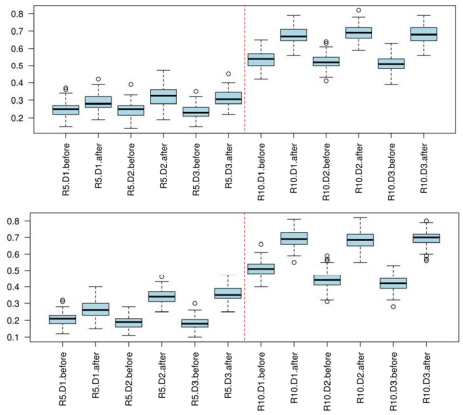

Fig. 5. Sensitivity of DEG prediction. Rn stands for *n* samples (replicates). Dm stands for *m* units of outlier distance. "before" and "after" stand for before and after outlier detection/correction. The upper panel shows the sensitivity measures for the first data set, which is composed of 10% outlier genes. The lower panel shows the sensitivity measures for the second data set, which is composed of 30% outlier genes. The critical *p* value was 0.05.

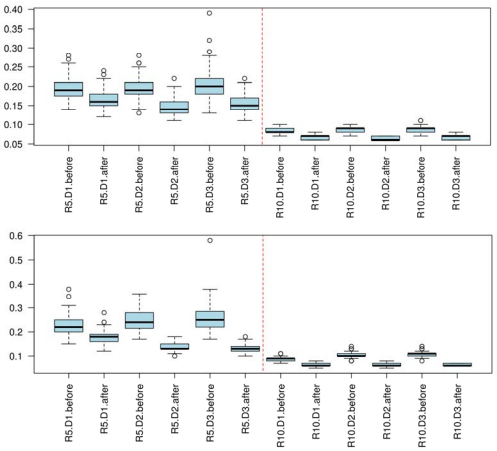

Fig. 6. pFDR measures for two data sets. The upper panel is for 10% outlier gene insertion and the lower panel is for 30% outlier gene insertion. Vertical axes stand for pFDR.

The next thing which is important is whether outlier genes can be well identified. This depends on how outliers were present in data. If an outlier sample has a small distance with non-outlier samples, the detection should not be very easy. When this distance was large, the detection should be easy. Fig. 7 shows one such simulation result, where the replicate number was five and the outlier distance varied from one to three. It can be seen that POD's performance on outlier gene prediction was negatively proportional to the outlier distance. However, all GOD algorithms failed to work reasonably. Among them, only LSOSS slightly outperformed other four GOD algorithms. Fig. 8 shows the AUR measures for two data sets with replicate number as five. It also shows that POD much outperformed other five GOD algorithms.

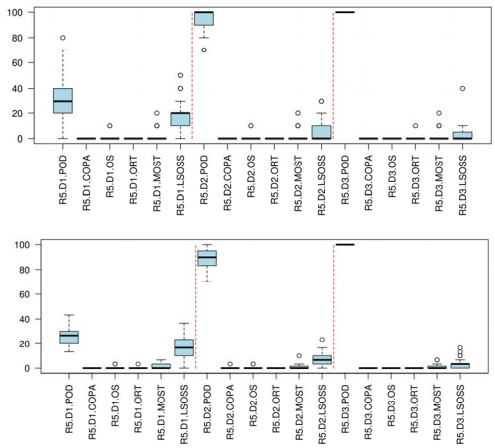

Fig. 7. The sensitivity of detecting outlier genes for data set with five replicates and 10% (upper) and 30% (lower) outlier gene insertion. The outlier distance varied from one to three. All six algorithms were compared. The sensitivity was scaled to percentage.

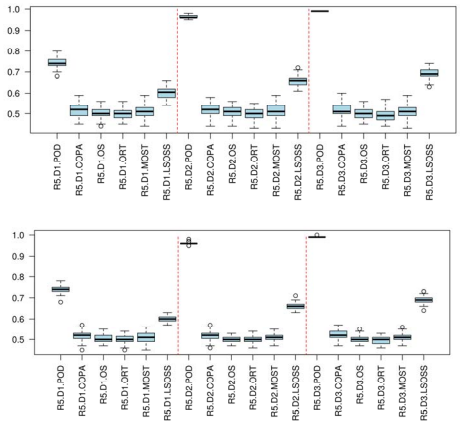

Fig. 8. AUR of detecting outlier genes for data set with five replicates and 10% (upper) and 30% (lower) outlier gene insertion. The outlier distance varied from one to three.

Based on the above analyses on the simulated data sets, we can see that population-based outlier detection can provide better outlier gene prediction accuracy. With a simple imputation approach for outlier correction, the DEG prediction accuracy can also be improved. With this confidence, we now analyse some cancer microarray gene expression data to examine how outlier genes distribute in data and to examine whether outlier correction (using simple imputation approach) can improve pFDR measurements. We also compare POD against five GOD algorithms.

Table 2. Results for the seven cancer data sets. q_0 and q_1 are the minimum *q* values before outlier correction. $pFDR_0$ and $pFDR_1$ are the corresponding maximum pFDR values.

	q_0	q_1	$pFDR_0$	$pFDR_1$
gds1439-prostate	0.02	1.2E-5	7.26E-06	2.09E-06
gse12630-breast	0.35	1.6E-4	3.12E-05	4.07E-06
gse12630-liver	0.99	0.0098	5.27E-05	9.50E-06
gse7410-cervical	0.05	1E-11	1.06E-05	2.01E-06
IGC-Breast	0.85	0.0006	1.69E-05	1.81E-06
IGC-Colon	0.13	2.7E-16	1.14E-05	1.39E-06
IGC-Sarcoma	0.99	0.0026	3.69E-05	2.42E-06

Table 2 summarises the results for the seven cancer data sets. Prior to outlier detection/correction, the minimum q value exceeded 0.01 in all data sets. After outlier detection/correction, none shows minimum q value larger than 0.01. Outlier detection/correction thus significantly reduced the minimum q value. Importantly, pFDR has been reduced significantly (about one magnitude lower) after using POD.

We now compare POD against GOD algorithms. First, we calculated the maximum gap between consecutive expressions for each gene. The control and test expressions were separately treated. For each gene, we would have two maximum gap values. Among them, we used the larger one for further analysis. For each gene, we also have six p values acquired from six algorithm including POD. We would like to examine how p values correlate with these maximum gaps. It was expected that if an outlier detection algorithm works well, the p values should more correlate with the maximum gaps. Table 3 shows the correlation measures between p values and maximum gaps for six algorithms and seven data sets. It can be seen that none of GOD algorithms shows good correlation with maximum gaps while POD does.

Table 3. Correlation between p values of outlier gene detection and maximum gaps between consecutive expressions.

	POD	COPA	OS	ORT	MOST	LSOSS
gds1439	0.81	0.060	0.09	0.02	0.060	0.033
gse12630	0.67	0.032	0.23	0.12	0.091	0.104
gse12630	0.71	-0.085	0.21	0.03	-0.05	-0.08
gse7410	0.61	0.128	0.29	0.21	0.173	0.208
IGC(B)	0.20	0.128	0.31	0.27	0.018	-0.18
IGC(C)	0.63	-0.047	0.17	0.08	0.044	0.030
IGC(S)	0.34	-0.080	0.07	-0.16	0.003	0.067

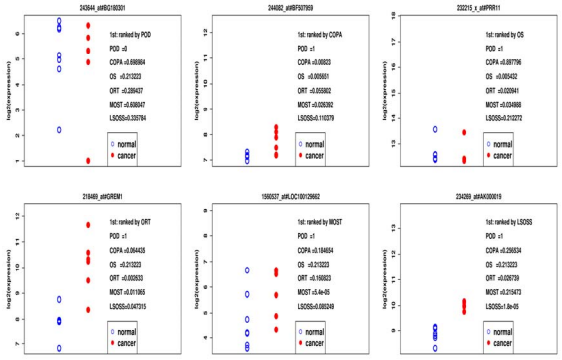

Fig. 9. The top outlier gene detected by six algorithms for data set GDS1439. Open dots stand for samples of normal patients. Filled dots stand for samples of cancer patients. The top captions are probe set IDs and gene symbols (separated by the hash key). The first line of texts indicates which algorithm is used to rank this gene as the top outlier gene. The following lines give the p values of six algorithms.

We examined whether an outlier gene detected by different algorithms does show significant separation between outliers and non-outliers. We collected top outlier gene detected by six algorithms based on the smallest p values. We plotted them in the same scale so as to compare different algorithms. Fig. 9 shows the top outlier gene detected by six algorithms for the data set GDS1439. Outlier genes ranked top by GOD algorithms normally did not show clearly outlier pattern compared with POD's prediction. For instance, gene 244082_at(BF507959) ranked top by COPA did not show a good outlier pattern. The p value of POD for this gene was 1. Gene 232215_x_at(PRR11) detected by OS did show outlier pattern as well. The outlier distance was not sufficiently large enough for POD to agree. Others also show same scenario.

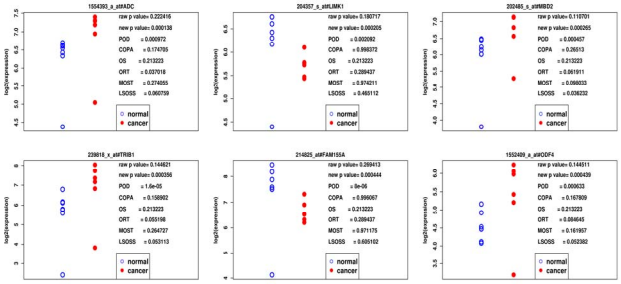

Fig. 10. Top six turnover genes for GDS1439.

We also examined the turnover genes (from non-DEGs to DEGs) after outlier detection/correction. A turnover gene is such a gene that it was identified as a non-DEG, but was identified as a DEG after outlier detection/correction using POD. We also examined how five GOD algorithms dealt with them, i.e. whether they were able to recognize the outlier characteristic of these genes. Fig. 9 shows the top six turnover genes for GDS1439 data set. The selection was based on q values. It can be seen why the correction of outliers using POD can switch a non-DEG to a DEG, i.e. generating a turnover gene. For instance, both control (normal) samples and test (cancer) samples of the gene ADC (probe set ID is 1554393_a_at) contained an outlier. POD's p value of being an outlier gene for this gene is 0.000972. The plot shows a clear pattern that the differentiation between normal and cancer samples was weakened by these two outliers. They were far below the non-outlier samples. These significant deviations largely enlarged the variances (or pooled variance) used for significance analysis using t test or modified t test. Because of significantly enlarged pooled variance, the calculated t statistic from t test or modified t test was small and the corresponding p value was large (0.222416). After outlier gene detection/correction, the new p value was 0.000138. Only COPA was able to identify this outlier gene and ORT marginally identified this outlier gene. Other three failed to identify this outlier gene. Other top five genes shown in Fig. 10 show the same pattern of turnover genes.

Fig. 11 illustrates the presence of potential NSOs in ten outlier genes with the smallest q values in the prostate cancer data set (GDS1439). These ten genes show significant up regulation which is apparent only after the removal of NSOs among the cancer samples. These genes have been studied in relation to prostate cancer [23-27]. Gene RAB11FIP4 has been found to be up-regulated in an analysis of microRNA transcriptome for prostate cancers based on 904 miRNAs [28]. NPY has been found to play a key role for prostate cancer progression [29] and NPY receptors are highly expressed in the androgen-independent prostate cancer cell lines contributing to cell proliferation [30]. ZNF595 has been found to be over-expressed in prostate cancer tumor [31].

IV. CONCLUSION

Outliers in microarray gene expression data can be classified into structural outliers and non-structural outliers (NSOs). NSOs are randomly distributed across all arrays and genes and are uninformative for gene expression differentiation. We showed that the presence of NSOs leads to severe under-estimation of gene expression differentiation and thus reduction in prediction power. NSOs cannot be efficiently detected using existing gene-specific outlier detection algorithms. We therefore propose a new algorithm for detecting NSOs by modelling the pooled sorted consecutive differential expressions as exponentially distributed. Imputation of detected NSOs can then lead to significant improvement in the quality of gene selection/ranking. We have illustrated this algorithm using simulated data sets and seven cancer data sets. In particular, we give some likely examples of overlooked significant genes due to NSOs.

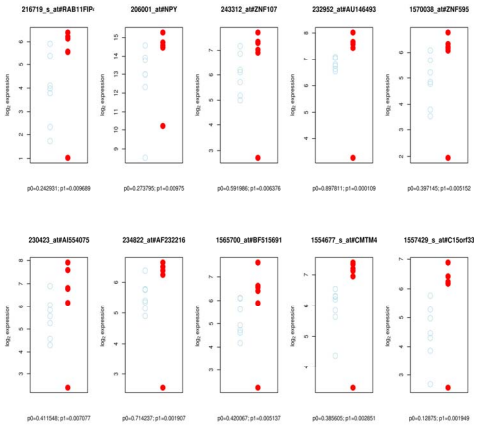

Fig. 11. $\log_2$ expressions of the top ten genes for the data set GDS1439-prostate. The open circles represent non-cancer samples and the filled circles represent cancer samples. p0 and p1 are the *t* test *p* values before and after outlier removal respectively.

REFERENCES

[1] A. Kauffmann, Huber, W, "Microarray data quality control improves the detection of differentially expressed genes," *Genomics,* vol. 95, pp. 138-42, 2010.

[2] R. Jaksik, Polanska, J, Herok, R, Rzeszowska-Wolny, J, "Calculation of reliable transcript levels of annotated genes on the basis of multiple probe-sets in Affymetrix microarrays," *Acta Biochim Pol,* vol. 56, pp. 271-7, 2009.

[3] G. Jurman, Merler, S, Barla, A, Paoli, S, Galea, A, Furlanello, C, "Algebraic stability indicators for ranked lists in molecular profiling," *Bioinformatics,* vol. 24, pp. 258-64, 2008.

[4] A. Shieh, Hung, YS, "Detecting outlier samples in microarray data," *Stat Appl Genet Mol Biol,* vol. 8, pp. 13, 2009.

[5] T. Ideker, Thorsson, V, Siegel, AF, Hood, LE, "Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data," *J Comput Biol,* vol. 7, pp. 805-17, 2000.

[6] I. Lonnstedt, Speed, T.P., "Replicated microarray data," *Statistica Sinica,* vol. 12, pp. 31-46, 2002.

[7] G. Tseng, Oh, MK, Rohlin, L, Liao, JC, Wong, WH, "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects," *Nucleic Acids Res,* vol. 29, pp. 2549-57, 2001.

[8] Z. Jia, Rahmatpanah, FB, Chen, X, Lernhardt, W, Wang, Y, Xia, XQ, Sawyers, A, Sutton, M, McClelland, M, Mercola, D, "Expression changes in the stroma of prostate cancer predict subsequent relapse," *PLoS One,* vol. 7, pp. e41371, 2012.

[9] B. Perez-Villamil, Romera-Lopez, A, Hernandez-Prieto, S, Lopez-Campos, G, Calles, A, Lopez-Asenjo, JA, Sanz-Ortega, J, Fernandez-Perez, C, Sastre, J, Alfonso, R, Caldes, T, Martin-Sanchez, F, Diaz-Rubio, E, "Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior," *BMC Cancer,* vol. 12, pp. 260, 2012.

[10] A. Reid, Attard, G, Brewer, D, Miranda, S, Riisnaes, R, Clark, J, Hylands, L, Merson, S, Vergis, R, Jameson, C, Høyer, S, Sørenson, KD, Borre, M, Jones, C, de Bono, JS, Cooper, CS, "Novel, gross chromosomal alterations involving PTEN cooperate with allelic loss in prostate cancer," *Mod Pathol,* vol. 25, pp. 902-10, 2012.

[11] H. Seol, Lee, HJ, Choi, Y, Lee, HE, kim, YJ, Kim. JH, Kang. E, Kim. SW, Park. SY, "Intratumoral heterogeneity of HER2 gene amplification in breast cancer: its clinicopathological significance," *Mod Pathol,* vol. 25, pp. 938-48, 2012.

[12] S. A. Tomlins, Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J.E., Shah, R.B., Pienta, K.J., Rubin, M.A., Chinnaiyan, A.M., "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," *Science,* vol. 310, pp. 644-8, 2005.

[13] R. Tibshirani, Hastie, T, "Outlier sums for differential gene expression analysis," *Biostatistics,* vol. 8, pp. 2-8, 2007.

[14] B. Wu, "Cancer outlier differential gene expression detection," *Biostatistics,* vol. 8, pp. 566-75, 2007.

[15] H. Lian, "MOST: detecting cancer differential gene expression," *Biostatistics,* vol. 9, pp. 411-8, 2008.

[16] B. Efron, Tibshirani, R, Storey, JD, Tusher, V, "Empirical Bayes analysis of a microarray experiment," *Journal of American Statistical Association,* vol. 96, pp. 1151-60, 2001.

[17] Y. Benjamini, Yekutieli, D., "The control of the false discovery rate in multiple testing under dependency," *The Annuals of Statistics,* vol. 29, pp. 1165-88, 2001.

[18] J. D. Storey, "A direct approach to false discovery rates," *J. R. Stat. Soc.,* vol. 64, pp. 479-98, 2002.

[19] J. Li, Paramita, P, Choi, KP, Karuturi, RK, "ConReg-R: Extrapolative recalibration of the empirical distribution of p-values to improve false discovery rate estimates," *Biol Direct,* vol. 6, pp. 27, 2011.

[20] W. J. Krzanowski, Hand, D.J., *ROC curves for continuous data*: CRC Press, 2009.

[21] C. E. Metz, "Basic principles of ROC analysis. ," *Seminars in Nuclear Medicine,* vol. 8, pp. 283-288, 1978.

[22] Y. Wang, Rekaya, R., "LSOSS: Detection of Cancer Outlier Differential Gene Expression," *Biomark Insights,* vol. 5, pp. 69-78, 2010.

[23] M. Stein, Dong, J, Wandinger-Ness, A, "Rab proteins and endocytic trafficking: potential targets for therapeutic intervention," *Adv Drug Deliv Rev,* vol. 55, pp. 1421-37, 2003.

[24] G. Mills, Jurisica, I, Yarden, Y, Norman, JC, "Genomic amplicons target vesicle recycling in breast cancer," *J Clin Invest,* vol. 119, pp. 2123-7, 2009.

[25] Y. Kim, Wuchty, S, Przytycka, TM, "Identifying causal genes and dysregulated pathways in complex diseases," *PLoS Comput Biol,* vol. 7, pp. e1001095, 2011.

[26] Y. Yao, Yang, WM, "Beyond histone and deacetylase: an overview of cytoplasmic histone deacetylases and their nonhistone substrates," *J Biomed Biotechnol,* vol. 2011, pp. 146493, 2011.

[27] K. Hemminki, Li, X, Sundquist, J, Sundquist, K, "Cancer risks in Crohn disease patients," *Ann Oncol,* vol. 20, pp. 574-80, 2009.

[28] D. Hebb, *The organization of behaviour*: John Wiley and Sons Inc, 1949.

[29] M. Bessarabova, Kirillov, E, Shi, W, Bugrim, A, Nikolsky, Y, Nikolskaya, T, "Bimodal gene expression patterns in breast cancer," *BMC Genomics,* vol. 11, no. S8, 2010.

[30] E. Fredlund, Staaf, J, Rantala, JK, Kallioniemi, O, Borg, A, Ringnér, M, "The gene expression landscape of breast cancer is shaped by tumor protein p53 status and epithelial-mesenchymal transition," *Breast Cancer Res,* vol. 14, pp. R113, 2012.

[31] F. Fitch, "Review: Warren S. McCulloch and Walter Pitts, A logic calculus of the ideas immanent in nervous activity," *Journal Symbolic Logic,* vol. 9, pp. 49-50, 1944.