Rank-Based Multiple Testing for Detecting Differentially Expressed Genes with Brief Overview of Parametric Methods

Bo Li Hossein Mansouri*

Abstract

Two major issues of concern in microarray data analysis are violation of the assumption of normality and influence of outliers. Since usually a very large number of tests are simultaneously carried out on microarray data, a serious concern is to control the familywise error rate (FWER), otherwise researchers may wrongly claim dozens even hundreds of genes to be differentially expressed. Another difficulty to deal with is deriving the theoretical distribution of the test statistic and calculation of the p-value. Resampling techniques such as permutation method and bootstrapping are used in data analysis to achieve better approximation of the p-values. This article provides a brief review of some permutation and bootstrap methods for the analysis of differentially expressed genes with application to a real dataset.

Key Words: microarray, rank tests, bootstrap, multiple testing, SAM

1. Introduction

The advances in microarray technology have enabled researchers to measure the intensity of spotted cDNA's on arrays simultaneously for thousands of genes. There has also been a concerted effort to develop statistical methods to analyze the large data sets that result from the microarray experiments one aspect of which is to detect the differentially expressed genes based on the spot intensity readings. The intensity readings often tend to be non-normal and contain a large number of outliers. This could be problematic when classical parametric methods are used. Rank-based methods are therefore recommended as an alternative to analyze microarray gene expression data because of their robustness to the violation of distributional assumptions and influence of outliers, Li and Mansouri (2015).

A large number of tests are often involved in the analysis of microarray data in order to detect differentially expressed genes. As such, control of familywise error rate (FWER) becomes of primary importance, otherwise researchers may wrongly claim dozens or even hundreds of genes to be differentially expressed.

Important works in the analysis of differentially expressed genes include Dudoit et al. (2002) who proposed Welch's t-test Statistic Step-down Procedure. This procedure uses a permutation method at each step of the step-down procedure. This procedure controls FWER strongly. Tusher et al. (2001) proposed the method of significance analysis of microarray (SAM). This method is based on a modified pooled t-test statistic. A graphical method called SAM plot is constructed to identify the significantly expressed genes.

Kerr and Churchill (2000) developed the analysis of variance (ANOVA) models for microarray gene expressions that take the ancillary sources of variation into consideration. This allows researchers to develop simultaneous tests for gene expressions based on linear models, Hsu et al. (2006) and Li and Mansouri (2015).

In this article, we provide a brief overview of the simultaneous testing procedures mentioned above and apply them to a well-known dataset and in the process we demonstrate that not all of these methods identify exactly the same sets of genes as differentially expressed. Further investigations may be required to the reliability of each method.

*Corresponding author. Email: hossein.mansouri@ttu.edu; Texas Tech University, Department of Mathematics and Statistics, Lubbock, TX 79409-1042, USA

2. Methods of Simultaneous Testing for Gene Expression

In this section we provide a brief overview of the methods of simultaneous testing. The methods in subsections 2.1 and 2.2 are based on a factorial ANOVA model and the remaining methods are based on two-sample formulation of the data.

2.1 Simultaneous rank tests

Let Y_{ijklm} be the normalized observation of the $i-th$ array, $j-th$ dye, $k-th$ treatment, $l-th$ gene, and $m-th$ replicate for $i = 1, \cdots, a,\ j = 1, 2,\ k = 1, 2,\ l = 1, \cdots, g,\ m = 1, \cdots, n_l$. It is assumed that Y_{ijklm} follows a linear model. To detect whether the relative expression of each gene θ_l differs from 0 or not, the hypotheses are:

$$H_{0l} : \theta_l = 0\ vs.\ H_{1l} : \theta_l \neq 0,\ l = 1, \cdots, g. \tag{1}$$

Li and Mansouri (2015) proposed simultaneous aligned rank tests (ART) for identifying differentially expressed genes based on a linear model. Let R_{ijklm} be the rank of the $ijklm^{th}$ reduced model residual (aligned observations) among all such residuals, where the least square (LS) estimates of the main effects and interactions based on the reduced model are used to generate the residuals.

The ART simultaneous tests are formed by replacing Y_{ijklm} in the full model with R_{ijklm}. The ART test statistic is given by

$$t_{l,(ART)} = \frac{\hat{\theta}_{l,(ART)}}{\sqrt{\hat{Var}[\hat{\theta}_{l,(ART)}]}},\ \ l = 1, \cdots, g \tag{2}$$

Where $\hat{\theta}_{l,(ART)}$ and $\hat{Var}[\hat{\theta}_{l,(ART)}]$ are the aligned rank estimate of θ_l and the estimated variance respectively, see Li and Mansouri (2015) for details.

The $\alpha-$level simultaneous rank tests reject the null hypothesis H_{0l} if

$$\frac{|\hat{\theta}_{l,(ART)}|}{\sqrt{\hat{Var}[\hat{\theta}_{l,(ART)}]}} \geq q_{\alpha,(ART)},\ l = 1, \cdots, g, \tag{3}$$

where $q_{\alpha,(ART)}$ is the upper $\alpha - th$ quantile of the sampling distribution of the maximum modulus statistics $\max_{l=1,\cdots,g} |t_{l,(ART)}|$. Since the sampling distribution of $\max_{l=1,\cdots,g} |t_{l,(ART)}|$ is unknown, the residual bootstrap technique of Efron and Tibshirani (1993) is used to estimate $q_{\alpha,(ART)}$.

2.2 Simultaneous tests based on least square estimates

Note that if we replace the aligned ranks R_{ijklm} to the normalized observations y_{ijklm}, it becomes simultaneous tests based on LS estimates, which are proposed by Hsu et al. (2006). As a brief overview, the simultaneous tests are carried out based on t-test statistics $t_l = \frac{\hat{\theta}_l}{\sqrt{\hat{Var}[\hat{\theta}_l]}},\ \ l = 1, \cdots, g$. Where $\hat{\theta}_l$ and $\hat{Var}[\hat{\theta}_l]$ are the LS estimate of θ_l and the estimated variance, respectively, see Hsu et al. (2006) and Li and Mansouri (2015) for details. Hsu et al. (2006) recommended rejecting the null hypothesis H_{0l} when $|t_l| = \frac{|\hat{\theta}_l|}{\sqrt{\hat{Var}[\hat{\theta}_l]}} \geq q_\alpha,\ l = 1, \cdots, g$, where q_α is the upper α-th quantile of $\max_{l=1,\cdots,g} |t_l|$. If the underlying distribution is normal, the quantiles q_α can be obtained by Probmc() function in SAS. Otherwise the quantiles q_α are generated through residual bootstrap method of Efron and Tibshirani (1993).

2.3 Welch's t-test Statistic Step-down Procedure

This method is due to Dudoit et al. (2002). Let $\mathbf{Y}$ be the data matrix with l rows corresponding to the genes being studied and $n = n_1 + n_2$ columns corresponding to the n_1 normalized observations of treatment 1 and n_2 normalized observations of treatment 2. Let H_{0l} denote the null hypothesis of no association between the expression level of gene l and the treatment, $l = 1, \cdots, g$. Let y_{klm} be the normalized observation of $k-th$ treatment, $l-th$ gene, and $m-th$ replicate, $k = 1, 2; l = 1, \cdots, g; m = 1, \cdots, n_k$, the test statistic is

$$t_l = \frac{\bar{y}_{1l\cdot} - \bar{y}_{2l\cdot}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The following permutation method is used to estimate the $p-value$ for each gene. For the $b-th$ iteration, b=1,...,B:

1. Permute the n columns of the data matrix $\mathbf{Y}$. The first (last) $n_1(n_2)$ columns now refer to the "fake" treatment 1 (treatment 2) group.

2. Compute the t-statistics $t_1(b), \cdots, t_g(b)$.

The unadjusted permutation $p-values$ are obtained by: $p_l = \frac{\sum_{b=1}^{B} I(|t_l(b)|>|t_l|)}{B}, l = 1, \cdots, g$.

Westfall and Young step-down adjusted $p-value$ (Westfall and Young, 1993) is given by $\tilde{p}_{r_l} = max_{l'=1,\cdots,l}(Pr(max_{l^* \in \{r_{l'}, \cdots, r_g\}} |T_{l^*}| \geq |t_{r_{l'}}| | H_0^C),\ l = 1, \cdots, g$ where H_0^C is the complete null hypotheses and $|t_{r_{l'}}|$ is the $l'-th$ largest absolute value of test statistics among $|t_{l'}|,\ l' = 1, \cdots, l$.

For the $b-th$ permutation, $b = 1, \cdots, B$,

1. Permute the n columns of data matrix $\mathbf{Y}$.

2. Compute the test statistics $t_1(b), \cdots, t_g(b)$.

3. Compute successive maxima of the test statistics: $u_g(b) = |t_{r_g}(b)|$,
$u_l(b) = max(u_{l+1}(b), |t_{r_l}(b)|), l = 1, \cdots, g-1$, where $|t_{r_l}(b)|$ is the $l-th$ largest absolute value of test statistics.

4. The adjusted $p-value$ is estimated by: $\tilde{p}_{r_l} = \frac{\sum_{b=1}^{B} I(u_l(b)>|t_{r_l}|)}{B}$ with monotonicity enforced by setting $\tilde{p}_{r_1} \leftarrow \tilde{p}_{r_1}, \tilde{p}_{r_l} \leftarrow max(\tilde{p}_{r_l}, \tilde{p}_{r_{l-1}}), l = 2, \cdots, g$.

2.4 Significance Analysis of Microarray (SAM)

This method is proposed by Tusher et al. (2001) uses a modified pooled t-test statistic:

$$t_l = \frac{\bar{y}_{1l\cdot} - \bar{y}_{2l\cdot}}{S_p(l)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} + S_0}$$

where $S_p(l) = \sqrt{\frac{1}{n_1+n_2-2}(\sum_{m=1}^{n_1}(y_{1lm} - \bar{y}_{1l\cdot})^2 + \sum_{m=1}^{n_2}(y_{2lm} - \bar{y}_{2l\cdot})^2)}$; the coefficient of variation of t_l was computed as a function of $S_p(l)$ in moving windows across the data. The value for S_0 was chosen to minimize the coefficient of variation; for details of the computation see SAM "Significance Analysis of Microarrays" users guide and technical document (Tibshirani et al., 2011).

Plot the observed values of t_l versus the estimated expected value $E(t_l)$. The estimates are obtained by the method of permutation. A regression line is fitted to the plot and the confidence bands are constructed by setting a tuning parameter Δ which is chosen according to the desired false discovery rate (FDR). FDR is estimated as follows: The

estimated number of falsely significant genes is the average of the number of genes called significant from all permutations. For example, the permuted data sets generated an average of 7.4 falsely significant genes, compared with 48 genes called significant, yielding an estimated FDR of 0.15.

2.5 Benjamini and Hochberg's step-up procedure

Benjamini and Hochberg (1995) proposed step-up testing procedures to control FDR. Let $p_{r_1} \leq p_{r_2} \leq ... \leq p_{r_g}$ be the ordered unadjusted $p - values$. The BH step-up adjusted $p - values$ are

$$\tilde{p}_{r_l} = \min_{l \leq l' \leq g} \{min(\frac{g}{l'} p_{r_{l'}}, 1)\}, l = 1, ...g.$$

3. Analysis of a real dataset

To illustrate the methods reviewed in the preceding section, we analyze the microarray data of liver gene expressions in "NZO/HILt mice" treated with $0.001\% CL316,243$ as compared with the control. This data is from The Jackson Laboratory (http://churchill.jax.org). The mRNA samples were extracted from mice liver cells. Dye-swap design is used in this experiment, Kerr et al. (2000). Using a graphical analysis of the normalized data set, Li and Mansouri (2015) demonstrated that the observations violate the assumption of normality in addition to the presence of a large amount of outliers.

For detection of differentially expressed genes in "NZO/HILt mice" data, the nominal significance level is set to $\alpha = 0.05$. Using Welch's t-test statistic step-down procedure of section 2.3, 4 genes were found significantly expressed based on 10,000 permutations. Using the unadjusted permutation $p-values$ coupled with BH step-up procedure of section 2.5, 6 genes were found significantly expressed. By using SAM of section 2.4, 7 genes were found significantly expressed. Using the parametric bootstrap simultaneous tests (BST) of section 2.2 with bootstrap size $B = 1,000$, 4 genes were found to be significantly expressed. Finally, using the ART method of section 2.1 with bootstrap size $B = 1,000$, 7 genes were found to be differentially expressed. The results are summarized in Table 1. An important point to note is that these methods identify different sets of genes as differentially expressed. However, all methods identified four genes with Clone ID 1, 46, 94, and 97.

Table 1: Significantly Expressed Genes (1: significant; 0: insignificant)

Clone ID	Welch's t step-down	SAM	BH step-up	BST	ART
1	1	1	1	1	1
18	0	1	1	0	1
2	0	1	0	0	1
46	1	1	1	1	1
88	0	1	1	0	0
94	1	1	1	1	1
97	1	1	1	1	1
98	0	0	0	0	1
Total	4	7	6	4	7

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 (Statistical Methodology)*, 57, 289–300.

Dudoit S., Yang Y.H., Callow M.J., and Speed T.P. (2002), "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, 12, 111–139.

Efron B., and Tibshirani R.J. (1993), *An introduction to the bootstrap* New York: Chapman and Hall.

Hsu J.C., Chang J.Y., and Wang T. (2006), "Simultaneous confidence intervals for differential gene expressions," *Journal of Statistical Planning and Inference*, 136(7), 2182–2196.

Kerr M.K., Martin M., and Churchill G.A. (2000), "Analysis of variance for gene expression microarray data," *Journal of Computational Biology*, 7(6), 819–837.

Li B., and Mansouri H.G. (2015), "Simultaneous rank tests for detecting differentially expressed genes," *Journal of Statistical Computation and Simulation*, DOI:10.1080/00949655.2015.1046073.

Tibshirani R., Chu G., Narasimhan B., and Li J. (2011), "samr: SAM: Significance Analysis of Microarrays," R package version 2.0. http://CRAN.Rproject. org/package=samr.

Tusher V.G., Tibshirani R., and Chu G. (2001), "Significance analysis of microarrays applied to the ionizing radiation response," *PNAS*, 98(9), 5116-5121.