A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data

Ronghui Xu* and Xiaochun Li

Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, USA

Received on October 22, 2002; revised on January 21, 2003; accepted on January 25, 2003

ABSTRACT

Motivation: In analyses of microarray data with a design of different biological conditions, ranking genes by their differential ‘importance’ is often desired so that biologists can focus research on a small subset of genes that are most likely related to the experiment conditions. Permutation methods are often recommended and used, in place of their parametric counterparts, due to the small sample sizes of microarray experiments and possible non-normality of the data. The recommendations, however, are based on classical knowledge in the hypothesis test setting.

Results: We explore the relationship between hypothesis testing and gene ranking. We indicate that the permutation method does not provide a metric for the distance between two underlying distributions. In our simulation studies permutation methods tend to be equally or less accurate than parametric methods in ranking genes. This is partially due to the discreteness of the permutation distributions, as well as the non-metric property. In data analysis the variability in ranking genes can be assessed by bootstrap. It turns out that the variability is much lower for permutation than parametric methods, which agrees with the known robustness of permutation methods to individual outliers in the data.

Contact: rxu@jimmy.harvard.edu

INTRODUCTION

In today’s molecular biology research, microarrays are rapidly becoming common laboratory tools. Often after some filtering of the genes, in order to identify specific genes of interest, a biologist would like to know: what are the most relevant \( k \) genes whose expressions are different under the different biological conditions? The number \( k \) here might depend on the feasibility of the biologist to search the literature and to do subsequent validation experiments in a reasonable time frame. This type of expression data analysis relates closely to statistical hypothesis testing, and has been considered by various authors (see for example, Tusher et al., 2001; Li and Wong, 2002; Kerr et al., 2002; Tsodikov et al., 2002). In the simple setting of two different biological conditions and replicates of gene expression chips under each condition, for example, the classical \( t \)-statistic or its variation (Tusher et al., 2001) can be used. However, due to the cost of microarray chips the number of replicates are usually small, and when the expression data or their transformations cannot be assumed to distribute normally, permutation methods are often used. Our work was motivated by our experience where, for selecting differentially expressed genes, parametric and permutation methods tend to agree to a fairly large extent. Surprisingly, however, little comparison has been done of the two types of approaches, either in the general statistical literature or with the specific applications to microarray expression data.

The particular gene expression data that motivated this work came from a biological experiment to understand the mechanisms by which integrins prevent apoptosis in leukemia B cells (Astier et al., 2003). After being starved without serum for 3 h, the cells were stimulated by either fibronectin, which rescues the cells from death, or poly-L-lysine (PLL), a control that allows the cells to stick but does not induce any specific signaling within the cells. The stimulation by either ligand was given for three different lengths of period of time, 1, 3 and 8 h. Microarray gene expression data were collected for each treatment at the end of each length of stimulation, plus time 0 when there was no treatment. Two precursor B leukemia cell lines were used, REH and Nalm-6. Each experiment was repeated once. This gives a total of 28 microarray chips (Affymetrix). From the empirical plots of the temporal gene expression profiles at the four time points, 0, 1, 3 and 8 h (data not shown), it appeared natural to model the expression data as a quadratic function of time. Therefore after processing with dChip (Li and Wong, 2001) and some preliminary filtering of the genes,

*To whom correspondence should be addressed.
we used a multiple linear regression model containing treatment (fibronectin or PLL), cell line, time, time$^2$, and their interactions. In ranking the genes according to their differential expressions under stimulation by fibronectin or PLL, we applied both parametric and permutation approaches. In the parametric approach, the $F$-statistic for treatment effect, including its interactions with time and time$^2$, for each gene was referred to the $F$-distribution with the corresponding number of degrees of freedom. Notice that this was the same distribution for all the genes considered, therefore it was equivalent to ranking the genes according to the $F$-statistics directly. In the permutation approach, the stimulant fibronectin or PLL was randomly relabelled, and a permutation $p$-value was obtained for each gene from 500 permutations. If we use a somewhat arbitrary cutoff of 0.1 for the $p$-values, the parametric method gives a set of top 72 genes, while the permutation method gives a set of top 64 genes. It is interesting that out of these two sets of top genes, 46 are common. See Astier et al. (2003) for more details on the genes and the validation results.

In this paper we aim to compare the performance of parametric versus permutation methods as they are applied to differential gene expression analysis. The goal is to rank the genes using one of the two approaches, as we did for the leukemia cell data. We first carry out such comparison under simulation, to assess the accuracy of the rankings using either method. We then compare the two methods on our leukemia cell data, using bootstrap to estimate the variability in the gene rankings that we provided to our biologists.

**SIMULATION**

**A single gene: hypothesis testing**

As mentioned above, our purpose is to rank the genes based on their expression data, using either parametric or permutation method. This is not exactly hypothesis testing, but it is related to the performance of the two types of methods when used for testing hypotheses. Although it is known that results from the two types of tests are similar for large samples (Cox and Hinkley, 1974; Good, 1994), little has been done in the literature for small samples in which we are interested.

Here we simulate two biological conditions, each with four samples. We assume equal variances of the two samples. Without loss of generality we test the hypotheses

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_a : \mu_1 < \mu_2.$$  

We shall perform one-sided test with $\alpha = 0.05$ (a two-sided test is basically two one-sided tests each at level $\alpha/2$). For the parametric $t$-test at level $\alpha$, we reject the null whenever the $t$-statistic $T_{\text{obs}} < t_{\alpha}(6) = -1.94$. For the permutation $t$-test at the same level, we reject the null whenever $T_{\text{obs}}$ is less than the $[\alpha m]$th ordered $t$-statistic from all the permutations of a given data set, where $m$ is the number of permutations and $[\cdot]$ indicates the integer part of a number. Here $m = \binom{4}{2} = 70$, and the actual type I error of the permutation test is $[0.05 \times 70]/70 = 0.043$. Notice that the actual type I error of the $t$-test equals to the nominal 0.05 if the data are normally distributed, but otherwise may change with the underlying distribution.

We simulate data from three different distributions: normal, lognormal and mixture of normals. The mixture of normals distribution is $0.3 \Phi((x - \mu)/\sigma) + 0.7 \Phi((x-\mu)/\sigma - 4)$, where $\Phi(\cdot)$ refers to the standard normal distribution. For each of the normal distributions involved (including in lognormal), the variance is 1 for both groups, while the mean is 0 under $H_0$ for both groups, and 0 and 1 for each of the two groups, respectively, under $H_1$.

Table 1 gives the simulation results. As we have already mentioned, the actual type I error of the $t$-test could differ from the nominal 0.05 if the data are not normally distributed. To make more ‘fair’ a comparison, we also estimated the power of the two tests when the actual $\alpha$’s are the same, at 0.043 level. From the table we see that the permutation test is more powerful when the underlying data are not normally distributed, especially when the actual $\alpha$’s are the same. When the data are normal, the performances of the two tests are comparable if we take into the account the variability associated with simulations. The standard error of the power estimate is 0.015 for 1000 simulations.

**Ranking genes: measures of expression differences**

As we have seen in the analysis of leukemia cell data described in Section Introduction a more direct use of either the parametric or the permutation method is to rank genes according to their differential expression values under different conditions. Here instead of hypothesis testing, we can view either of the two approaches as providing a measure of distance between the gene expressions under different conditions. To formalize the idea, suppose that there are two different conditions.

**Table 1. A simulated comparison of the two test statistics**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Test</th>
<th>Realized $\alpha$</th>
<th>Power</th>
<th>Power$^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$t$-test</td>
<td>0.05</td>
<td>0.35</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Permutation</td>
<td>0.043</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>Log-normal</td>
<td>$t$-test</td>
<td>0.047</td>
<td>0.24</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>Permutation</td>
<td>0.043</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>Mixture of normals</td>
<td>$t$-test</td>
<td>0.065</td>
<td>0.38</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>Permutation</td>
<td>0.043</td>
<td>0.36</td>
<td>0.36</td>
</tr>
</tbody>
</table>

*Power is the probability of rejection at $\alpha = 0.043$. Number of simulations $B = 1000$. 

Notice that the actual type I error of the $t$-test at level $\alpha$ is basically two one-sided tests each at level $\alpha/2$. For the parametric $t$-test at level $\alpha$, we reject the null whenever the $t$-statistic $T_{\text{obs}} < t_{\alpha}(6) = -1.94$. For the permutation $t$-test at the same level, we reject the null whenever $T_{\text{obs}}$ is less than the $[\alpha m]$th ordered $t$-statistic from all the permutations of a given data set, where $m$ is the number of permutations and $[\cdot]$ indicates the integer part of a number. Here $m = \binom{4}{2} = 70$, and the actual type I error of the permutation test is $[0.05 \times 70]/70 = 0.043$. Notice that the actual type I error of the $t$-test equals to the nominal 0.05 if the data are normally distributed, but otherwise may change with the underlying distribution.

We simulate data from three different distributions: normal, lognormal and mixture of normals. The mixture of normals distribution is $0.3 \Phi((x - \mu)/\sigma) + 0.7 \Phi((x-\mu)/\sigma - 4)$, where $\Phi(\cdot)$ refers to the standard normal distribution. For each of the normal distributions involved (including in lognormal), the variance is 1 for both groups, while the mean is 0 under $H_0$ for both groups, and 0 and 1 for each of the two groups, respectively, under $H_1$.

Table 1 gives the simulation results. As we have already mentioned, the actual type I error of the $t$-test could differ from the nominal 0.05 if the data are not normally distributed. To make more ‘fair’ a comparison, we also estimated the power of the two tests when the actual $\alpha$’s are the same, at 0.043 level. From the table we see that the permutation test is more powerful when the underlying data are not normally distributed, especially when the actual $\alpha$’s are the same. When the data are normal, the performances of the two tests are comparable if we take into the account the variability associated with simulations. The standard error of the power estimate is 0.015 for 1000 simulations.

**Ranking genes: measures of expression differences**

As we have seen in the analysis of leukemia cell data described in Section Introduction a more direct use of either the parametric or the permutation method is to rank genes according to their differential expression values under different conditions. Here instead of hypothesis testing, we can view either of the two approaches as providing a measure of distance between the gene expressions under different conditions. To formalize the idea, suppose that there are two different conditions.
We can write the microarray gene expression data as \( \{X_i\}_{i=1}^n \) and \( \{Y_j\}_{j=1}^m \) under the first and the second condition, respectively, where \( n \) and \( m \) are the sample sizes, \( X_i = (x_{i1}, \ldots, x_{ik})' \), \( Y_j = (y_{j1}, \ldots, y_{jk})' \), and \( K \) is the total number of genes under consideration. Let the \( X \) and \( Y \) sample follow multivariate distributions \( G_1 \) and \( G_2 \), respectively. The genes are ranked according to \( \text{dist}(G_{1k}, G_{2k}) \), where \( G_{1k} \) and \( G_{2k} \) are the \( k \)th marginal distribution of \( G_1 \) and \( G_2 \), respectively. For example, the top number one gene is \( \text{argmax}_k \{ \text{dist}(G_{1k}, G_{2k}) \} \). The distance here, corresponding to the statistic used, is symmetric but may not satisfy the triangle inequality, so it is not necessarily a metric. The expression data for gene \( k \) is \( \{X_{ik}\}_{i=1}^n \) and \( \{Y_{jk}\}_{j=1}^m \), and it is not difficult to see that either the parametric \( t \)-statistic (squared) or its permutation \( p \)-value provides a finite sample estimated distance between the two marginal distributions.

In the following we compare the parametric statistic with its permutation version. To keep the comparison straightforward and focused, we restrict ourselves to squared Euclidean distances here, and do not consider other types such as ordinary (i.e. non-squared) Euclidean distance (Mielke and Berry, 2001), or rank-based methods.

**Two conditions** We first simulate 500 genes under two different conditions, each with four replicated expression values as in the previous setting. Although in practice the expression values from some genes are likely correlated, as explained above in ranking the genes we only make use of the marginal distributions of each gene. Therefore the simulated expression data for different genes are independent. The data are again simulated under normal, lognormal and mixture of normal distributions, with the mean expression of gene \( k \) in the second group being higher than that in the first group by \( k \times C \), \( k = 1, \ldots, 500 \), \( C = 4 \) for lognormal and \( C = 1 \) for normal and mixture of normals. The error variance is 100 for normal and mixture of normals, and 0.5 on the log scale for lognormal. The genes are ordered randomly before being ranked using one of the methods. Figures 1–3 give the results for five simulations each. Similar types of plots have also been used in Newton et al. (2000) and Tsodikov et al. (2002).

In the figures the horizontal axis indicates the number, \( x \), of top genes up to 300 that have been found using one of the methods, and the vertical axis is the number of true top \( x \) genes among them. The 45° line then corresponds to knowing the truth, and any other method would generally give a curve that lies under the 45° line. The dashed lines correspond to ranking the genes randomly for each of the five simulations (the bottom group of lines on all three figures). The solid curves are obtained by using the squared \( t \)-statistic to rank the genes directly, and the dotted lines are obtained by using the permutation \( p \)-values of the \( t \)-statistic to rank the genes. As expected, using the \( t \)-statistic directly performs the best when the expression data are normally distributed. The permutation method does not out-perform the parametric method in all three cases; this may appear surprising when compared with the results of the previous subsection. However, notice that there are only 70 total possible permutations for two samples with size 4 each, which gives 70 possible different \( p \)-values. Using these \( p \)-values to rank 500 genes, we are likely to encounter many ties. In particular, for a given gene, when the two samples do not overlap, it can be shown that the permutation \( p \)-value of the \( t \)-statistic is always 1/70, regardless of the magnitude of difference in the means. This shows that the distance function underlying the permutation method is not a metric, and is...
SS is a monotonically decreasing function of $\sigma$ for any permutation of the data value. So the permutation the other hand, the distance function underlying the statistic under the normal setting is proportional to $(\mu + \tau)/\sigma^2$. We observe that, $F$ is a monotonically decreasing function of $SS_{within}$, and for any permutation of the data $SS_{total}$ doesn’t change. Therefore for the non-overlapping case mentioned before, any permutation other than the original grouping of the two samples will increase $SS_{within}$ and decrease the $F$-value. So the permutation $p$-value is always 1/70. On the other hand, the distance function underlying the $t$-statistic under the normal setting is proportional to $(\mu_1 - \mu_2)/\sigma$, and is a strictly monotone function of the group differences in units of standard deviations. For lognormal with group means $\mu_i$ ($i = 1, 2$) and equal variances $\sigma^2$ on the logscale, the lognormal distributions themselves have means $\exp(\mu_i + \sigma^2/2)$ and variances $\exp(2\mu_i + \sigma^2)[\exp(\sigma^2) - 1]$. The expectation of the $t$-statistic is no longer a monotone function of a location shift; in fact this is not a simple location-shift model as the variances are no longer equal in the two groups. This case also shows us that it is important to choose the right scales (transformations) for the data before ranking. In our simulation the parametric and permutation methods had similar performances for both the lognormal and the mixture models.

**Temporal expression data** Next we simulate temporal gene expression data, in a way similar to the leukemia cell data. Notice that different batches of cells were treated for different lengths of time, therefore the expression data at different time points are independent. We model four time points just as in the leukemia cell data: 0, 1, 3 and 8 h. At each time point and for each of the two treatments we have four replicates. We generate data from the quadratic model

$$y \sim trt + t + t^2 + trt \times t + trt \times t^2.$$ 

The three regression parameters affecting the underlying differential expressions are: treatment effect and treatment by time interactions. To order the genes in such a three-dimensional parameter space, for our simulation we consider the simple scenario of simultaneously increasing the magnitude of these three parameters. For the permutation method, treatment labels are permuted within each time point, therefore we have a total of $4^3 = 70^3$ possible permutations. We randomly select 500 permutations to compute the permutation $p$-values. We again investigate the performances under three error structures, normal, lognormal and mixture of normals.

From Figures 4–6 we notice that once again the permutation method does not outperform the parametric $F$-statistic. We also noticed many ties in the permutation $p$-values, even for 500 permutations. Although we may advocate to choose as many permutations as one can afford, the gain appears limited once a reasonable number of permutations have been carried out.

**LEUKEMIA CELL DATA**

Unlike the previous section where we know the truth under simulation settings, in this section we use bootstrap to
assess the variability in ranking genes due to the small sample size of the leukemia cell expression data. The variability is an aspect not examined in the simulation studies. Our data can be seen as a random sample from an underlying data-generating mechanism, and the question is: if we draw another random sample, how many top (say 100) genes will be rediscovered by either the parametric or permutation method?

In bootstrapping the expression data, we randomly resampled the 28 chips with replacement. Figure 7 shows the result of five bootstraps. Figure 7a is similar to the figures of the last section, except that instead of comparison against true ranking, we plot the agreement between one bootstrap ranking and each of the rest four bootstrap rankings. Figure 7b shows the average of the four agreements and the pointwise 95% confidence intervals using the parametric and permutation method, respectively. Notice that the confidence intervals can be narrowed simply at the computational cost of more bootstrap runs. On average, using the parametric method, it is estimated that about 10 out of the top 100 genes will be rediscovered in another random sample. Since the average curve is close to being linear, we may say that about 10% of top \( x \leq 100 \) genes can be rediscovered in another sample; we call this percentage the rediscovery rate, which in general depends on \( x \).

For the permutation method, the rediscovery rate for top 100 genes is estimated to be much higher, at about 53%. This can at least be partially explained by the discreteness of the permutation distribution. For example, if we take a closer look at bootstraps 1 and 2, there are 55 overlapping top 100 genes; out of these 55 genes, there are only 11 unique permutation \( p \)-values. Therefore the discreteness of the permutation method, while it affects the accuracy as shown in the simulations, also effectively reduces the variation in ranking genes.

**DISCUSSION**

In this paper we compared the performance of parametric versus permutation methods as they are applied to microarray gene expression data. The conventional wisdom is that when data are not normally distributed, the permutation method is better than the parametric methods which are based on the normality assumption. Our simulation re-illustrated this in the setting of hypothesis testing.
Ranking genes, as already described earlier (Ranking genes: measures of expression difference), is different in that a statistic is used as a distance function. When using the permutation method, due to the discreteness and the resulting violation of strict monotonicity, we are likely to have ties, especially if the expression values are highly differential and when the noise level is relatively low. This can reduce both the accuracy and the variability of the permutation method in ranking genes.

Our experience from simulation as well as the leukemia data indicate that parametric method has reasonable performance. In light of its low cost in computation, we recommend that it always be used. Following that if there are reasonable number of replicates such as the temporal leukemia data so that ties are not expected to be a problem, then permutation ranking should also be carried out. The leukemia investigators took the common top genes which were found by both methods; this might be a worthy approach. In addition we recommend to use non-specific filters to weed out genes under-expressed and lack of variation to narrow down the list of candidates as much as possible. Equally important is the choice of the scale of data for ranking genes.

For the leukemia cell data, where the true ranking of the genes is unknown, we used bootstrap to assess the variability in rankings by both methods. In light of the small sample sizes of microarray experiments, variability is an important issue to be addressed. The amount of variability can change from experiment to experiment, depending on the underlying data-generating mechanism that may not be verified. It is therefore helpful to assess the variability on a case by case basis. The concept of rediscovery rate complements the aspect that is described by, for example, the false discovery rate (FDR, Benjamini and Hochberg, 1995; Efron et al., 2001; Storey, 2001). The FDR is given in the framework of multiple testing, which certainly is the case for microarray gene expression data, especially if we want to emphasize a pre-defined significance of the discovered genes. The rediscovery rate, on the other hand, is associated with ranking, usually without a cutoff as in defining a significance level. In practice, with sufficiently large number of bootstraps, we would be able to estimate for each gene, what the rediscovery rate is. Although this may still be hindered by today’s computational power, especially if permutations are used, it would nonetheless provide very useful information to a biologist in his or her decision concerning avenues of further investigation.

We would also like to note that the temporal structure of the leukemia cell data has generality because treatments are often given for different lengths of time in biology laboratories. The type of polynomial modelling seems to capture the treatment effects over time. Fine tuning, nonetheless, will probably be needed to make the method more tailored towards microarray gene expression data.

Finally, although the leukemia cell data was generated using Affymetrix technology, our findings are not confined to Affymetrix and also applies to cDNA dual-channel (red–green) arrays, once the arrays are normalized and log-ratios of intensities have been computed.

ACKNOWLEDGEMENTS

The authors would like to thank Anne L. Astier and Marek Svoboda for their collaboration and feedback on the leukemia cell data analysis. We would also like to thank the editors and the referees for their helpful comments.

REFERENCES


