The t-mixture model approach for detecting differentially expressed genes in microarrays

Shunpu Zhang

Available at: http://works.bepress.com/shunpu_zhang/3/
The $t$-mixture model approach for detecting differentially expressed genes in microarrays

Shuo Jiao · Shunpu Zhang

Abstract The finite mixture model approach has attracted much attention in analyzing microarray data due to its robustness to the excessive variability which is common in the microarray data. Pan (2003) proposed to use the normal mixture model method (MMM) to estimate the distribution of a test statistic and its null distribution. However, considering the fact that the test statistic is often of $t$-type, our studies find that the rejection region from MMM is often significantly larger than the correct rejection region, resulting an inflated type I error. This motivates us to propose the $t$-mixture model (TMM) approach. In this paper, we demonstrate that TMM provides significantly more accurate control of the probability of making type I errors (hence of the familywise error rate) than MMM. Finally, TMM is applied to the well-known leukemia data of Golub et al. (1999). The results are compared with those obtained from MMM.

Keywords $t$-mixture · Normal mixture · Permutation

Introduction

The use of microarray technology makes it possible to monitor the expression levels of thousands of genes simultaneously. A common goal of analyzing the genomewide expression data generated from this technology is to detect genes with differential expression under two conditions. Now, as the cost of microarray experiments keeps decreasing, replicated microarray experiments are feasible. The replicated measurements of expression levels form the basis of the methods in our paper.

In recent years, numerous nonparametric approaches for detecting significantly differentially expressed (DE) genes have been proposed in the literature (Efron et al. 2001; Tusher et al. 2001; Pan et al. 2003; Zhang, 2006), among others. In these nonparametric methods, the null distribution (the distribution of the test statistic for equivalently expressed (EE) genes) is estimated directly from the repeated measurements of gene expression levels under each condition.

In the mixture model method (MMM; Pan et al. 2003), finite normal mixture models are used to estimate the distribution of the test statistic and the null distribution. However, noticing the fact that both the test statistics are usually heavy-tailed in practice, it is more natural to view them as the observations from a mixture of the $t$-distributions. As pointed out in McLachlan and Peel (2000), the estimates of the component means and variances can be affected by observations that are atypical of the components in a finite normal mixture model. As a result, MMM may underfit the true underlying densities and produce critical values too small in absolute values. If the significance level approach is used, this will produce inflated type I error rates and lead to inflated familywise error rate (FWER), and if the $p$ value approach is used, this will significantly underestimate the true $p$ values and leads to more false rejections.

To avoid the underfit problem of MMM, some alternatives have been proposed in the literature (Allison et al. 2002; McLachlan et al. 2005, 2006). The $t$-mixture model (TMM) approach has been proposed as a heavy-tailed alternative to the normal mixture model in the software...
EMMIX-GENE developed by McLachlan et al. (2002) for clustering the microarray-expressed data. However, the use of TMM for the detection of DE genes was not discussed in their paper. In our paper, we propose to use the TMM approach to estimate the distributions of the test statistic and its corresponding null distribution. Following the lines of Pan et al. (2003), the null distribution is estimated from the permuted sets of null scores. We will show that TMM can adapt to the atypical observations better than MMM and provide more accurate critical values. In addition, our simulations show that no obvious improvement can be made by applying TMM on more than one permuted set of null scores. Finally, we further illustrate the difference between TMM and MMM by applying them to the leukemia data of Golub et al. (1999).

Materials and methods

The test statistic and null statistic

Suppose that $Y_{ij}$ is the expression level of gene $i$ in array $j$ ($i=1, 2, \ldots, n; j=1, \ldots, j_1, j_1+1, \ldots, j_1+j_2$), and the first $j_1$ and last $j_2$ arrays are obtained under the two different conditions. A general statistical model is:

$$Y_{ij} = a_i + b_jx_i + \varepsilon_{ij}$$  \hspace{1cm} (1)

where $x_i=1$ for $j\leq j_1$ and $x_i=0$ for $j> j_1$. To determine whether the mean expression levels under the two conditions are the same is equivalent to testing the following hypothesis: $H_0: b_1=0$ against $H_1: b_1 \neq 0$.

The standard two sample $t$-statistic for testing this hypothesis is:

$$Z_i = \frac{\bar{Y}_{i(1)}-\bar{Y}_{i(2)}}{s_{i(1)}^2/j_1 + s_{i(2)}^2/j_2}$$  \hspace{1cm} (2)

where $\bar{Y}_{i(1)}, \bar{Y}_{i(2)},$ and $s_{i(1)}^2, s_{i(2)}^2$ are the sample means and sample variances of the $Y_{ij}$'s under two conditions, respectively. Under the normality assumption on $Y_{ij}$, the null distribution of $Z_i$ is approximately $t$-distributed.

However, when the normality assumption is violated, the use of the $t$-distribution is not appropriate. A class of nonparametric statistical methods has been proposed to overcome this problem. The basic idea of the nonparametric methods is to estimate the null distribution of the test statistic $Z$ by treating the values of the test statistic, when being applied to the permuted microarray data, as the true null scores one would expect from EE genes. However, recent research reveals that such practice is problematic. Zhao and Pan (2003) showed that one needs to modify the test statistic $Z$ and construct its corresponding null statistic such that the null statistic, when being applied to the permuted microarray data, provides the correct null scores. Several methods for constructing the test statistic and the null statistic were proposed in Zhao and Pan (2003). However, it was pointed out in Pan (2003) that the methods of Zhao and Pan (2003) are quite restrictive. For example, it requires even number of observation under each experimental condition. Improvements over Zhao and Pan (2003) were made in Pan (2003) in which he proposed the following test statistic and its corresponding null statistic:

$$Z_1 = \frac{\bar{Y}_{11}+\bar{Y}_{12} - \bar{Y}_{21}+\bar{Y}_{22}}{\sqrt{\frac{s_{11}^2/j_1+s_{12}^2/j_2}{4} + \frac{s_{21}^2/j_1+s_{22}^2/j_2}{4}}}$$  \hspace{1cm} (3)

$$Z_1 = \frac{\bar{Y}_{11}-\bar{Y}_{12} + \bar{Y}_{21} - \bar{Y}_{22}}{\sqrt{\frac{s_{11}^2/j_1+s_{12}^2/j_2}{4} + \frac{s_{21}^2/j_1+s_{22}^2/j_2}{4}}}$$  \hspace{1cm} (4)

where $\bar{Y}_{mn}$ and $s_{mn}^2$ ($m,n=1, 2$) are the sample means and sample variances of the subsamples obtained by partitioning the sample within each condition, respectively. By noticing the fact that the observations under the same condition are from the same population, Zhang (2006) provided an improved version of the above test statistic and null statistic:

$$Z_1 = \frac{\bar{Y}_{11}+\bar{Y}_{12} - \bar{Y}_{21}+\bar{Y}_{22}}{\sqrt{\frac{1}{j_1+1/j_1} \frac{s_{11}^2+1}{4} + \frac{1}{j_2+1/j_2} \frac{s_{22}^2}{4}}}$$  \hspace{1cm} (5)

$$z_1 = \frac{\bar{Y}_{11}-\bar{Y}_{12} + \bar{Y}_{21} - \bar{Y}_{22}}{\sqrt{\frac{1}{j_1+1/j_1} \frac{s_{11}^2+1}{4} + \frac{1}{j_2+1/j_2} \frac{s_{22}^2}{4}}}$$  \hspace{1cm} (6)

where $\bar{Y}_{mn}$ ($m,n=1,2$) are defined the same as those in Eqs. 3 and 4, and $s_{11}^2, s_{22}^2$ are the pooled sample variances of the observations under each condition, respectively. More details about the construction of Eqs. 3, 4, 5 and 6 can be found in Pan (2003) and Zhang (2006).

The $t$-mixture model

In MMM, Pan et al. (2003) used a normal MMM to estimate the density functions of $Z_1$ and $z_1$ defined by Eqs. 3 and 4 denoted by $f$ and $f_0$, respectively. As mentioned in the “Introduction” section, it is more reasonable to view the test and null statistics as the observations from a TMM. In the TMM, it is assumed that the data are from several components with distinct $t$-distributions. That is, both $f$ and $f_0$
are considered to be a mixture of the t-distributions with probability density function:

\[ h(z; \psi_g) = \sum_{i=1}^{g} \pi_i \varphi(z; \mu_i, \Sigma_i, v_i) \]  

(7)

where \( \varphi(z; \mu_i, \Sigma_i, v_i) \) denotes the density of the t-distribution with mean \( \mu_i \), variance \( \Sigma_i \), and degree of freedom \( v_i \). The coefficients \( \pi_i \)'s are the mixing proportions and \( g \) is the number of components, which can be selected adaptively. The vector \( \psi_g \) denotes all the unknown parameters \( (\pi_i, \mu_i, \Sigma_i, v_i)_{i=1, \ldots, g} \) in Eq. 7. The TMM is fitted by the maximum likelihood method using an expectation conditional maximization (ECM) algorithm (Liu and Rubin 1995).

In the ECM algorithm, \( \psi_g \) is partitioned as \( (\varphi_1^T, \psi_2^T, \ldots, \psi_g^T) \) with \( \varphi_1 = (\pi_i, \mu_i, \Sigma_i | i = 1, \ldots, g) \) and \( \psi_2 = (v_i | i = 1, \ldots, g) \). Given the \( n \) \( p \)-dimensional observations \( y_j, j = 1, \ldots, n \), on the \( (k+1) \)th iteration of the ECM algorithm, the estimates of the parameters in Eq. 7 are updated in two steps:

**Step 1.**

\[ \pi_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)}}{n}, \]

\[ \mu_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} y_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}, \]

\[ \Sigma_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (y_j - \mu_i^{(k+1)}) (y_j - \mu_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \]

where

\[ \tau_{ij}^{(k)} = \frac{\pi_i^{(k)} \varphi(y_j; \mu_i^{(k)}, \Sigma_i^{(k)}, v_i^{(k)})}{\sum_{i=1}^{g} \pi_i^{(k)} \varphi(y_j; \mu_i^{(k)}, \Sigma_i^{(k)}, v_i^{(k)})} \]

and

\[ \mu_i^{(k)} = \frac{\pi_i^{(k)} \varphi(y_j; \mu_i^{(k)}, \Sigma_i^{(k)}, v_i^{(k)})}{\sum_{i=1}^{g} \pi_i^{(k)} \varphi(y_j; \mu_i^{(k)}, \Sigma_i^{(k)}, v_i^{(k)})}. \]

**Step 2.** Find \( v_i^{(k+1)} \) as a solution of the equation:

\[ -\text{diagramma} \left( \frac{\nu_i}{2} \right) + \log \left( \frac{\nu_i}{2} \right) + 1 + \frac{1}{\sum_{j=1}^{n} \tau_{ij}^{(k+1/2)}} \]

\[ \sum_{j=1}^{n} \tau_{ij}^{(k+1/2)} \left( \log \left( \nu_i^{(k+1/2)} \right) - u_{ij}^{(k+1/2)} \right) + \text{diagramma} \left( \frac{\nu_i^{(k)} + p}{2} \right) - \log \left( \frac{\nu_i^{(k)} + p}{2} \right) = 0 \]

where

\[ \tau_{ij}^{(k+1/2)} = \frac{\pi_i^{(k+1)} \varphi(y_j; \mu_i^{(k+1)}, \Sigma_i^{(k+1)}, v_i^{(k)}/2)}{\sum_{i=1}^{g} \pi_i^{(k+1)} \varphi(y_j; \mu_i^{(k+1)}, \Sigma_i^{(k+1)}, v_i^{(k)}/2)} \]

and

\[ u_{ij}^{(k+1/2)} = \frac{\nu_i^{(k)} + p}{\nu_i^{(k)} + p + \left( y_j - u_i^{(k+1)} \right)^T \Sigma_i^{(k+1)-1} \left( y_j - u_i^{(k+1)} \right)}. \]

At convergence, we obtain \( \psi_g^\infty \) as the maximum likelihood estimate. Because the ECM may reach local maxima instead of global maxima, it is desirable to run this algorithm multiple times with different initial values and choose the estimates with the largest likelihood. Another important issue with TMM is the determination of the number of components \( g \). In this paper, we use the Bayesian information criterion (BIC) as the model selection criterion:

\[ \text{BIC} = -2 \log \left( h(z; \psi_g) \right) + t_g \log(n) \]

where \( h(z; \psi_g) \) is defined in Eq. 7, and \( t_g \) is the number of independent parameters in the probability density function.

Due to the fact that the values of the test and null statistics in microarray analysis are usually heavy-tailed, we expect to see better performance of TMM than MMM when the given significance level is very small. We will verify this in the next section.

**Results**

**Simulated data**

**Simulation set-ups** To see the control of the type I error by MMM and TMM, we first consider the situation in which the null hypothesis holds. In this case, the expression levels of genes under the two experimental conditions are drawn from the same distribution. Two types of distributions are used: the standard normal and the t-distribution with \( df = 3 \), representing the most commonly used distribution and a heavy-tailed distribution, respectively. The sample sizes are \( j_1 = 4 \) and \( j_2 = 4 \) for each gene, reflecting small sample sizes common in many microarray experiments. Here are the steps of the simulations.

1. For each distribution, 10,000 genes are generated.
2. We estimate \( f_0 \) with both TMM and MMM. Only \( f_0 \) is considered in this paper because the simulated data under the two conditions are from the same distribution.
3. For each of the two estimates of $f_0$, a rejection region $z: |z| > z_0$ is established such that $P(|z| > z_0) = \alpha$ where $\alpha$ is the given significance level. In our study, $\alpha = .005, .001, .0005, .0001, .00001, .000005$. In this way, we can get $z_0 = z_{f_0}^T$ with $f_0$ estimated by TMM and $z_0 = z_{f_0}^M$ with $f_0$ estimated by MMM. At last, out of all genes, we counted the proportion of genes which have a corresponding $|Z_i| > z_0^T$ and the proportion of genes having a corresponding $|Z_i| > z_0^M$, ($i=1, \ldots, 10,000$), which are the type I error rates for TMM and MMM, respectively.

We repeated steps 1–3 using both $z^T$ (Pan 2003) and $z^M$ (Zhang 2006) for 100 times. Table 1 of the Electronic Supplementary Material summarizes the average type I error rates for TMM and MMM. We find that MMM gives severely inflated type I error rates compared to the specified $\alpha$ when $\alpha$ is small and TMM gives significantly more accurate estimates. However, for the set-up with the $t$-distributed data, we notice that when $\alpha$ is greater than 0.0001, TMM is outperformed by MMM, which motivates us to see how well TMM and MMM fit the null statistics by checking the QQ plot between them. In Fig. 1, we can see that within a certain distance from 0, TMM has a greater departure from the reference line comparing to MMM. This is why TMM gives higher false-positive rates and may even have larger variation for the false-positive rates than MMM. However, this problem of TMM is limited to the case when the data are from the $t$-distribution and when the level of significance is relatively high. Hence, it is usually not a problem for the analysis of microarray data. For example, if the genome-wide level of significance is chosen as 0.01, the gene-specific level for a microarray data of 5,000 genes from the Bonferroni correction is 0.01/5000 = 0.000002, which is much smaller than 0.0001, below which we found TMM underperforms MMM in our simulations. It can also be seen in Fig. 1 that, when it comes to the tail, TMM tends to stay closer to the reference line, and as the significance level decreases, its performance starts to improve and becomes significantly better than MMM.

Another important factor which may affect the performance of TMM is the stability of the estimates of the degree of freedoms. For TMM, we found that the estimates of the degree of freedom for the $t$-distributed data are not very stable, from 2.83 to 13.00. This is part of the reason why TMM loses to MMM when the levels of significance are relatively high.

Table 2 of the Electronic Supplementary Material is obtained under similar set-up as that of Table 1 of the Electronic Supplementary Material. The purpose is to compare the performance of TMM and MMM with respect to using only one permuted set of null scores and using all possible permuted sets of null scores (under the prescribed setup, there are in total nine distinct permuted sets available). We are also interested in the effect of the number of permutations in the presence of DE genes. For this purpose, we generated a total of 5,000 genes among which 200 were DE genes. The numbers of replicates under the two conditions are chosen as $j_1=4$ and $j_2=6$, respectively. For the first 100 DE genes, the data under condition 1 are generated from $N(0,1)$ and the data under condition 2

<table>
<thead>
<tr>
<th>Model</th>
<th>Genome-wide significance level</th>
<th>Total identified DE gene</th>
<th>Correctly identified gene</th>
<th>Correctly identified proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMM</td>
<td>0.05</td>
<td>130</td>
<td>42</td>
<td>0.3230</td>
</tr>
<tr>
<td>MMM</td>
<td>0.01</td>
<td>153</td>
<td>43</td>
<td>0.2810</td>
</tr>
<tr>
<td>TMM</td>
<td>0.005</td>
<td>77</td>
<td>37</td>
<td>0.4805</td>
</tr>
<tr>
<td>MMM</td>
<td>0.001</td>
<td>107</td>
<td>41</td>
<td>0.3832</td>
</tr>
<tr>
<td>TMM</td>
<td>0.0005</td>
<td>61</td>
<td>35</td>
<td>0.5738</td>
</tr>
<tr>
<td>MMM</td>
<td>0.001</td>
<td>85</td>
<td>37</td>
<td>0.4353</td>
</tr>
<tr>
<td>TMM</td>
<td>0.0005</td>
<td>37</td>
<td>27</td>
<td>0.7297</td>
</tr>
<tr>
<td>MMM</td>
<td>0.05</td>
<td>53</td>
<td>32</td>
<td>0.6038</td>
</tr>
<tr>
<td>TMM</td>
<td>0.0005</td>
<td>26</td>
<td>23</td>
<td>0.8846</td>
</tr>
<tr>
<td>MMM</td>
<td>0.01</td>
<td>47</td>
<td>31</td>
<td>0.6596</td>
</tr>
</tbody>
</table>

Fig. 1 Plot of the comparison between TMM and MMM

Table 1 Comparison of the results from TMM and MMM at given levels of significance for the leukemia data
are generated from $N(3,1)$. For the remaining 100 DE genes, the data under condition 1 were generated from $N(0,1)$ and the data under condition 2 were generated from $N(-1,1)$. The data for EE genes are generated from $N(0,1)$ under both conditions.

From Tables 2 and 3 of the Electronic Supplementary Material, we find that TMM estimates are not affected by the change of the number of permutations in all set-ups (normal data, $t$-distributed data, and the data with DE genes). In contrast, the performances of MMM seems to depend on the number of permutations: the more permuted sets of null scores are used, the better the estimates get. Due to these findings, we suggest using TMM with just one permutation in practice.

The leukemia data

The leukemia data of Golub et al. (1999) is one of the most studied gene expression data set. This data set includes 27 acute lymphoblastic leukemia (ALL) samples and 11 acute myeloid leukemia (AML) samples for 7,129 genes. The goal is to find genes with differential expression between ALL and AML. Based on biological justification, Thomas et al. (2001) analyzed this data set and identified 50 genes as the most expressed and related genes to the disease, including 25 most expressed genes for AML and 25 for ALL.

At each given genome-wide level of significance $\alpha$, we computed the cut-offs $z_T^{\alpha}$ for TMM and $z_M^{\alpha}$ for MMM. The Bonferroni method was used to adjust for multiplicity of the tests. Then, we calculated the test scores $Z_i$ of the leukemia data and found the genes with $Z_i > z_T^{\alpha}$ for TMM and $Z_i > z_M^{\alpha}$ for MMM. These genes are called the predicted DE or significant genes. Finally, we examined the predicted DE genes to see which method contains more genes from the Thomas et al. (2001) list of DE genes.

The results from the comparison are summarized in Table 1. When the genome-wide level $\alpha$ is 0.005, TMM correctly identifies 35 out of 61 (53.38%) DE genes from the list while MMM correctly identifies 37 out of 85 (38.32%) DE genes. Table 1 shows that TMM consistently has a greater proportion of correctly identified genes than MMM, which means that TMM always has a smaller false-positive rate, regardless of the levels of significance. Table 2 contains a list of the DE genes identified by both TMM and MMM at $\alpha=0.0005$. The $p$ values from Thomas et al. (2001) are given as the reference. As we expected, MMM always gives smaller $p$ values than TMM does. In other
words, MMM tends to provide smaller p values due to its incapability to capture the true variability in the data and hence may contain more false-positive rates than TMM at the same level of significance.

Discussion

We have proposed to use the TMM for detecting the DE genes in a microarray experiment. It is demonstrated that TMM can provide more accurate control of the probability of type I error than MMM. Because the main focus of this paper is to introduce the TMM approach, we only discussed the control of false-positive rates by controlling the FWER. FWER can only provide control of the false-positive rates when no genes under consideration are DE. Hence, such control only works fine when there are none or very few genes which are actually DE among all the genes in consideration. In the situations that a relatively substantial amount of genes are DE, more efficient control of the false-positive rates can be achieved by controlling the false discovery rate (FDR) (Benjamini and Hochberg 1995; Storey and Tibshirani 2003).

Another point we want to stress is our proposal to only use one set of the permuted null scores when using TMM. The current practice of the permutation-based methods often suggests using all possible permutations (or a subset of it if the total number of available permutations is too large). Such suggestions ignore the possible pitfalls which the correlated sets of permuted null scores could cause when using a method such as the EM algorithm (Dempster et al. 1977) which requires i.i.d. observations. We believe that this proposal is important because not only can it significantly save computational time, which is the major concern of the finite mixture model approach, it can also avoid the problems caused by the use of the correlated permuted sets of null scores.

Acknowledgements The authors would like to thank the referee and the editor for their valuable comments which have significantly improved the quality of the paper.

Reference