Cancer Outlier Analysis Based on Mixture Modeling of Gene Expression Data

Keita Mori, Tomonori Oura, Hisashi Noma, and Shigeyuki Matsui

1 Department of Statistical Science, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
2 Clinical Trial Coordination Office, Shizuoka Cancer Center, 1007 Shimonagakubo, Nagaizumi-cho Sunto-gun, Shizuoka 411-8777, Japan
3 Asia-Pacific Statistical Sciences, Lilly Research Laboratories Development Center of Excellence Asia Pacific, Eli Lilly Japan K. K. Sannomiya Plaza Building 7-1-5 Isogamidori, Chuo-ku, Kobe, Hyogo 651-0086, Japan
4 Department of Data Science, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

Correspondence should be addressed to Keita Mori; kmori@ism.ac.jp

Received 30 January 2013; Accepted 23 March 2013

Academic Editor: Shinto Eguchi

Copyright © 2013 Keita Mori et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Molecular heterogeneity of cancer, partially caused by various chromosomal aberrations or gene mutations, can yield substantial heterogeneity in gene expression profile in cancer samples. To detect cancer-related genes which are active only in a subset of cancer samples or cancer outliers, several methods have been proposed in the context of multiple testing. Such cancer outlier analyses will generally suffer from a serious lack of power, compared with the standard multiple testing setting where common activation of genes across all cancer samples is supposed. In this paper, we consider information sharing across genes and cancer samples, via a parametric normal mixture modeling of gene expression levels of cancer samples across genes after a standardization using the reference, normal sample data. A gene-based statistic for gene selection is developed on the basis of a posterior probability of cancer outlier for each cancer sample. Some efficiency improvement by using our method was demonstrated, even under settings with misspecified, heavy-tailed t-distributions. An application to a real dataset from hematologic malignancies is provided.

1. Introduction

Heterogeneity of the expression of oncogenes within the same histological cancers is considered to have significant implications for understanding disease biology, identifying risk groups, and optimizing patient treatment [1, 2]. Recently, Tomlins et al. [3] argued that traditional analytical methods, for example, a two-sample t-statistic, which search for common activation of genes across a class of cancer samples, will fail to detect cancer genes which show differential expression in a subset of cancer samples or cancer outliers. They developed the “cancer outlier profile analysis” (COPA) method to detect cancer genes with such heterogeneous expression profiles within cancer samples and revealed subtypes of prostate cancer patients defined by recurrent chromosomal aberration.

Inspired by the COPA statistic, some authors have proposed other methods for detecting cancer-related genes with cancer outlier profiles in the framework of multiple testing [4–6]. However, such cancer outlier analyses will generally suffer from a serious lack of power because the analysis attempts to detect relatively small fractions of cancer outliers; the signal contained in the data is relatively limited, compared with that in the standard multiple testing setting where common activation of cancer-related genes for all cancer samples is supposed. As information sharing across units in the data generally improves efficiency of the analysis, we propose a simple efficient method via information sharing across
both genes and cancer samples. Specifically, we propose a parametric normal mixture modeling of gene expression levels of cancer samples across genes after a standardization using the reference, normal sample data. Then, a gene-based statistic for gene selection is proposed on the basis of a posterior probability of cancer outlier for each cancer sample. This posterior probability itself is to provide a useful index to aid identifying cancer outliers for a selected gene.

This paper is organized as follows. After providing a brief summary of the existing multiple testing methods for the cancer outlier analysis in Section 2, we provide the proposed method in Section 3. We assess performance of our methods via simulations in Section 4. An application to a real dataset from hematologic malignancies is given in Section 5. Finally, concluding remarks appear in Section 6.

2. Existing Multiple Testing Methods for Cancer Outlier Analysis

We suppose a microarray study to detect cancer-related genes from a large pool of G genes based on their gene expression levels measured for n samples, comprised of n₀ samples from a normal class and n₁ samples from a cancer class. The gene expression data considered here comprise normalized log ratios from two-color cDNA arrays or normalized log signals from oligonucleotide arrays (e.g., Affymetrix GeneChip). For gene g (g = 1, . . . , G), let xgi be the expression value for sample i (i = 1, . . . , n₀) in the normal class and let ygj be that for sample j (j = 1, . . . , n₁) in the cancer class. The multiple testing methods developed for analyzing cancer outliers intend to a one-sided testing. Without loss of generality, we are interested in detection of activated genes that are overexpressed or upregulated in a subset of cancer samples, that is, cancer outliers. For detecting cancer-related genes with over- or underexpressions, one may perform two one-sided tests separately, one for detecting cancer-related genes with overexpressions and the other for detecting those with underexpressions.

The traditional two-sample t-statistic for gene g is defined as

\[
t_{g} = \frac{\bar{y}_{g} - \bar{x}_{g}}{s_{g}},
\]

where \(\bar{y}_{g}\) is the mean expression value in the cancer samples, \(\bar{x}_{g}\) is the mean expression value in the normal samples, and \(s_{g}\) is the usual pooled standard error estimate for gene g (g = 1, . . . , G). The t-statistic is efficient in detecting cancer-related genes on which most cancer samples are activated, but may not be efficient for those with cancer outlier profiles.

Tomlins et al. [3] defines the COPA statistic as

\[
\text{Copa}_{g} = \frac{q_{g}(y_{gj} : 1 \leq j \leq n_{1}) - \text{med}_{g}}{\text{mad}_{g}},
\]

where \(q_{g}(\cdot)\) is the rth percentile of the expression level, \(\text{med}_{g}\) is the median of expression values, and \(\text{mad}_{g}\) is the median absolute deviation of expression values in all of the samples:

\[
\text{med}_{g} = \text{median}(x_{gi}, y_{gj} : i = 1, \ldots, n_{0}, \ j = 1, \ldots, n_{1}),
\]

\[
\text{mad}_{g} = 1.4826 \times \text{median}\left(\left|x_{gi} - \text{med}_{g}\right|, \left|y_{gj} - \text{med}_{g}\right|\right),
\]

\[
i = 1, \ldots, n_{0}, \ j = 1, \ldots, n_{1}.
\]

(3)

The value of r in \(q_{g}(\cdot)\), which represents a threshold in determining cancer outlier, is specified by the user, such as r = 75, 90, or 95.

Instead of using a fixed r percentile value, approximately equivalent to using the information from only one sample, the use of additional outlier samples can be more efficient. Specifically, the OS statistic [4] is defined as

\[
\text{OS}_{g} = \frac{\sum_{i \in R_{g}} (y_{gj} - \text{med}_{g,i})}{\text{mad}_{g}}.
\]

Here the set of cancer outliers, \(R_{g}\), is heuristically identified by \(R_{g} = \{j : y_{gj} > q_{75}(x_{gi}, y_{gj} : i = 1, \ldots, n_{0}, \ j = 1, \ldots, n_{1}) + \text{IQR}(x_{gi}, y_{gj} : i = 1, \ldots, n_{0}, \ j = 1, \ldots, n_{1})\}\), where IQR(D) is the interquartile range of the data D, IQR(D) = \(q_{75}(D) - q_{25}(D)\).

Wu [5] proposed the ORT statistic through identifying cancer outliers relative to the normal sample, rather than the pooled sample. Specifically, the ORT statistic is defined as

\[
\text{ORT}_{g} = \frac{\sum_{i \in O_{g}} (y_{gj} - \text{med}_{g,i})}{\text{mad}_{g}^{*}},
\]

where \(O_{g} = \{j : y_{gj} > q_{75}(x_{gi} : i = 1, \ldots, n_{0}) + \text{IQR}(x_{gi} : i = 1, \ldots, n_{0}), \text{med}_{g,x} = \text{median}(x_{gi} : i = 1, \ldots, n_{0}), \text{med}_{g,y} = \text{median}(y_{gj} : j = 1, \ldots, n_{1})\}\), and

\[
\text{mad}_{g}^{*} = 1.4826 \times \text{median}\left(\left|x_{gi} - \text{med}_{g,x}\right|, \left|y_{gj} - \text{med}_{g,y}\right|\right),
\]

\[
i = 1, \ldots, n_{0}, \ j = 1, \ldots, n_{1}.
\]

(6)

As the COPA, OS, and ORT statistics are criticized because the outliers are arbitrarily defined, Lian [6] considers all possible values of the outlier threshold. Specifically, for the ordered gene expressions for the cancer samples, \(y_{g1} \geq y_{g2} \geq \cdots \geq y_{g_{n_{1}}},\) the most statistic is defined as

\[
\text{MOST}_{g} = \max_{1 \leq k \leq n_{1}} \left\{ \frac{\sum_{j \leq k} (y_{gj} - \text{med}_{g,x}) / \text{mad}_{g}^{*} - \mu_{k}}{\sigma_{k}} \right\},
\]

(7)

where \(\mu_{k} = E[\sum_{j \leq k} z_{j}],\) and \(\sigma_{k}^{2} = \text{Var}[\sum_{j \leq k} z_{j}]\) for \(z_{1} > z_{2} > \cdots > z_{n_{1}},\) the order statistics of \(n_{1}\) samples from the standard normal distribution. The standardization in the parenthesis is to make different values of the statistic comparable for different values of the outlier threshold, \(k (k = 1, \ldots, n_{1}).\)
3. The Proposed Method

3.1. Mixture Modeling of Gene Expression Data. In order for information sharing across both genes and cancer samples, we propose a simple parametric normal mixture modeling of gene expression data of cancer samples. As the existing multiple testing methods, for each gene, we consider standardized gene expressions of the cancer samples based on the reference, normal sample data,

\[ u_{gj} = \frac{y_{gj} - \bar{X}_g}{s_{gx}}, \]

where \( s_{gx} \) is the usual standard error estimate within the normal samples for gene \( g \) (\( g = 1, \ldots, G; j = 1, \ldots, n_j \)). Again, the standardization intends to make all gene expression data from the cancer samples comparable across genes. We then assume the finite normal mixture model with the three components,

\[ f(u_{gj}) = \pi_0 f_0(u_{gj}) + \pi_1 f_1(u_{gj}) + \pi_2 f_2(u_{gj}). \]

The density function \( f_0 \) corresponds to the null component with no differential expressions for the reference, normal sample data. The densities \( f_1 \) and \( f_2 \) correspond to the nonnull components (i.e., cancer outliers) of underexpression and overexpression, respectively, for the normal sample data. We specify normal distributions, \( N(0, 1^2) \), \( N(\delta_1, 1^2) \), and \( N(\delta_2, 1^2) \), for \( f_0 \), \( f_1 \), and \( f_2 \), respectively, \( \pi_k \) represents the mixing proportion (\( k = 0, 1, 2 \)), and \( \pi_0 + \pi_1 + \pi_2 = 1 \). We denote \( Z_{gjk} \) as an unobservable indicator random variables, such that \( Z_{gjk} = 1 \) if the (standardized) expression level, \( u_{gj} \), of cancer sample \( j \) on gene \( g \) belongs to the \( k \)th component, and \( Z_{gjk} = 0 \) otherwise (\( g = 1, \ldots, G; j = 1, \ldots, n_j \)). We estimate the parameters, \( \delta_1 \), \( \delta_2 \), and \( \pi_k \)'s, via applying the EM algorithm to cope with the unobservable indicator variable \( Z_{gjk} \) in the mixture model (e.g., [7]).

3.2. Statistics for Gene Selection. The posterior probability, \( w_{gjk} \), that \( Z_{gjk} = 1 \), that is, the expression level \( u_{gj} \) belongs to the \( k \)th component, provides a basis for gene selection,

\[ w_{gjk} = \frac{\hat{\pi}_k f_k(u_{gj})}{\hat{f}(u_{gj})}. \]

For detecting overexpressed genes, possibly with a cancer outlier profile (as a one-sided testing), we propose to use the following gene-based statistic for gene selection,

\[ S_g = 1 - \frac{1}{n_1} \sum_{j=1}^{n_1} (1 - w_{gjj}). \]

This statistic may correspond to one minus the posterior probability that none of samples are cancer outliers with overexpressions. We will select genes with greatest values of \( S_g \). Gene-based statistics for detecting underexpressed cancer-related genes can be similarly developed.

In our framework, we can also derive a similar gene-based statistic for detecting under- or overexpressed genes (as a two-sided testing). One has

\[ T_g = 1 - \frac{1}{n_1} \sum_{j=1}^{n_1} \left( 1 - (w_{gjj} + w_{gjj}) \right). \]

It is important to note that the posterior probabilities, \( w_{gjk} \), themselves can serve as a helpful index to aid identifying cancer outlier samples for a particular (selected) gene. In contrast, the existing cancer outlier methods do not provide such an expression-level statistic for identifying cancer outlier samples.

Unlike the existing statistics for cancer outlier analysis, the statistic, \( S_g \), does not involve any particular cancer outlier threshold, so that cancer-related genes with various proportions of cancer outliers (\( \phi \) in Section 4), even those with common activation across all cancer samples, could be detected. However, as \( S_g \) is a composite of the posterior probabilities from all of the cancer samples, cancer-related genes with smaller proportions of cancer outlier will be more difficult to be detected because the statistic will be more dominated by the posterior probabilities from the cancer samples other than cancer outliers. The impact of the proportion of cancer outlier will be investigated in Section 4.

4. Simulation Study

We conducted a simulation study to assess the performance of our method in detecting cancer-related genes with cancer outlier profiles. We considered a microarray study with \( G = 10000 \) genes for \( n = 40, 80, \) or \( 200 \) samples, where the first half of samples were from the normal class and the latter half from the cancer class, that is, \( n_0 = n_1 = n/2 \). Of note, for a given \( n_1 \), the power of the analysis will improve as \( n_0 \) increases because more precise estimates of the mean and variance of the normal sample data become available in the standardization \( u_{gj} \) before fitting the mixture model (9) to detect cancer-related genes. We generated the gene expression levels for each gene from the standard normal distribution \( N(0, 1^2) \) or the central \( t \)-distribution with 20 degrees of freedom to assess the impact of deviation from the normality assumption. No interaction across genes was assumed. We supposed that \( G \) genes were divided into the three gene components according to the mixture model (9), that is, the null, underexpression, and overexpression component. The mixing proportions were set to as \( \pi_0 = 0.6 \), \( \pi_1 = \pi_2 = 0.2 \). For each nonnull gene with under- or overexpressions, the proportion of cancer outliers in the cancer samples, \( \phi \), was set to be \( \phi = 0.1, 0.3 \), or 0.5. We supposed a common difference or effect size in gene expression between the cancer outlier samples and the other samples (normal samples and nonoutlier cancer samples) across nonnull genes and set the value of \( \delta_1 \) as \(-2.0\) and that of \( \delta_2 \) as \(2.0\). For each configuration, we performed gene selection based on the \( t \)-statistic, COPA, OS, ORT, MOST, and the proposed one-sided statistics, \( S_g \), for detecting overexpressed cancer-related genes.
Figure 1: ROC curves that plot TPR versus FDR for normally distributed gene expression data.

We assessed the false discovery rate (FDR) and true positive rate (TPR), defined as the proportion of false positives in the set of significant genes and the proportion of selected true positives in the set of all of the overexpressed genes (= \( G r_{x} \)), respectively. Note that the TPR corresponds to average power in multiple testing (e.g., [8, 9]). We conducted 200 simulations to obtain average TPR for a given value of FDR for each method, as the estimates of TPR were highly stable for \( G = 10000 \) values of each statistic obtained in a single simulation run.

Figures 1 and 2 show ROC curves that plot the TPR and FDR for various numbers of significant genes in multiple testing for normally distributed and \( t \)-distributed gene expressions, respectively.

For normally distributed gene expressions, the gene selection based on the proposed statistic, \( S_{g} \), generally provided the greatest values of TPR (for a given value of FDR). As is expected, the proposed gene selection based on \( S_{g} \) provided greater TPR as \( \phi \) increased. The gene selection based on the \( t \)-statistic provided the smallest values of TPR, especially when the proportion of cancer outliers is small, such as \( \phi = 0.1 \), but the TPR improved for greater values of the proportion, such as \( \phi = 0.5 \), as is expected. The COPA and OS methods performed worst among the methods except the \( t \)-test, especially for greater values of \( \phi \), such as \( \phi = 0.5 \). In particular, the performance of the OS method was largely deteriorated for \( \phi = 0.5 \). The ORT and MOST methods generally provided comparable TPR values, but less than those of the proposed method based on \( S_{g} \).

For \( t \)-distributed gene expressions, similar trends were observed. Again, the proposed method based on \( S_{g} \) provided greatest TPR in general, except for the scenario with \( n = 40 \).
and $\phi = 0.1$, although the degree of its superiority to the other methods, such as the ORT and MOST methods, becomes smaller, compared with the settings with normally distributed gene expressions. The COPA and OS methods again provided the smallest values of TPR, especially when $\phi$ is large, such as $\phi = 0.5$.

5. Application

We illustrate how the proposed method can capture the heterogeneity of cancer samples through its application to a microarray gene expression data from the myelodysplastic syndromes (MDSs) [10]. The MDSs are complex hematologic malignancies with heterogeneous clinicopathological features with various chromosomal aberrations. In order to discover the heterogeneous clinicopathological features of MDSs, possibly including those related to prognosis, we adopted the proposal using mixture distributions method for 139 MDSs and 69 nonleukemias (samples from bone marrow mononuclear cells from nonleukemic conditions), which were regarded as cancer and normal samples, respectively. Here, following Mills et al. [10], we removed samples of the chronic myelomonocytic leukemia disease type from MDS samples. We first adopted the RMA normalization [11] to the raw data (Raw data (CEL files) downloaded from Gene Expression Omnibus database (GEO, http://www.ncbi.nlm.nih.gov/geo/, accession number GSE15061)). We make

**Figure 2:** ROC curves that plot TPR versus FDR for $t$-distributed gene expression data.
Table 1: The number of overlaps in selected genes between the gene selection methods in the example of hematologic malignancies. Top 200 genes were selected by each method.

<table>
<thead>
<tr>
<th>t-statistic</th>
<th>COPA</th>
<th>OS</th>
<th>ORT</th>
<th>MOST</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-statistic</td>
<td>—</td>
<td>13</td>
<td>14</td>
<td>50</td>
<td>56</td>
</tr>
<tr>
<td>COPA</td>
<td>13</td>
<td>—</td>
<td>150</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>OS</td>
<td>14</td>
<td>150</td>
<td>—</td>
<td>139</td>
<td>108</td>
</tr>
<tr>
<td>ORT</td>
<td>50</td>
<td>0</td>
<td>139</td>
<td>—</td>
<td>151</td>
</tr>
<tr>
<td>MOST</td>
<td>56</td>
<td>99</td>
<td>108</td>
<td>151</td>
<td>—</td>
</tr>
<tr>
<td>Proposed</td>
<td>56</td>
<td>51</td>
<td>86</td>
<td>89</td>
<td>75</td>
</tr>
</tbody>
</table>

The proposed statistic would therefore be effective for selecting cancer-related genes that are involved in relatively major activation among cancer samples. Modification of the statistic for selecting cancer-related genes with more minor activation (i.e., small $\phi$) is a subject for future research. Another important subject would be the addition of a gene-level mixture structure, that is, nonnull and null genes in terms of the association with cancer, to provide a more formal basis for evaluating false positives and true positives in gene selection.

In this paper, we have attempted to improve the efficiency of the cancer outlier analysis through information sharing across genes and cancer samples. In our simulations, the proposed gene selection based on a parametric normal mixture model of gene expression data demonstrated some improvement in efficiency for detecting cancer-related genes with moderate to large proportions of cancer outlier ($\phi \geq 0.3$), even under settings with heavy-tailed $t$-distributions. The proposed statistic would therefore be effective for selecting cancer-related genes that are involved in relatively major activation among cancer samples. Modification of the statistic for selecting cancer-related genes with more minor activation (i.e., small $\phi$) is a subject for future research. Another important subject would be the addition of a gene-level mixture structure, that is, nonnull and null genes in terms of
Figure 3: Histograms of the standardized expression values of three genes selected by our method, but not by the other methods.
Acknowledgment

This research was partly supported by a Grant-in-Aid for Scientific Research (24240042) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Submit your manuscripts at
http://www.hindawi.com