Biostatistics (2007), **8**, 1, *pp.* 2–8
doi:10.1093/biostatistics/kxl005
Advance Access publication on May 15, 2006

Outlier sums for differential gene expression analysis

ROBERT TIBSHIRANI*

Department of Health Research & Policy and Department of Statistics, Stanford University, Stanford, CA 94305, USA
tibs@stat.stanford.edu

TREVOR HASTIE

Department of Statistics and Department of Health Research & Policy, Stanford University, Stanford, CA 94305, USA

SUMMARY

We propose a method for detecting genes that, in a disease group, exhibit unusually high gene expression in some but not all samples. This can be particularly useful in cancer studies, where mutations that can amplify or turn off gene expression often occur in only a minority of samples. In real and simulated examples, the new method often exhibits lower false discovery rates than simple *t*-statistic thresholding. We also compare our approach to the recent cancer profile outlier analysis proposal of Tomlins *and others* (2005).

Keywords: Cancer; COPA; Gene expression analysis; Microarray.

1. INTRODUCTION

We consider methods for detecting differentially expressed genes in a set of microarray experiments. We consider the simple case of m genes measured across two experimental conditions. A number of authors have proposed methods for detecting differential gene expression; Dudoit *and others* (2002) and Allison *and others* (2006) give summaries.

One widely used approach to this problem is as follows. We compute a two-sample *t*-statistic T_i for each gene, and then call a gene significant if $|T_i|$ exceeds some threshold c. Various values of c are tried using permutations of the sample labels to estimate the false discovery rate (FDR) for the procedure for each c. A threshold c is finally chosen based on the estimates of FDR and other considerations, such as the ballpark number of significant genes that is desirable. This recipe roughly describes the strategy used, for example, in the significance analysis of microarrays (SAM) procedure (Tusher *and others*, 2001). The SAM procedure can be applied to other test statistics for a wide variety of data types, such as paired, censored, or time-course data.

In a study of mutations in prostate cancer, Tomlins *and others* (2005) introduced a method called "cancer profile outlier analysis" (COPA) for detecting what they call "oncogene outliers." These are genes which show a systematic increase in expression, but only for a small number of cancer samples. They show that COPA can be more powerful than the usual *t*-statistic in these cases. In related work, Lyons-Weiler

*To whom correspondence should be addressed.

and others (2004) proposed the permutation percentile separability test with a similar objective: to find genes that are overexpressed only in a subset of cases.

The COPA work inspired us to study this problem and look for better ways of detecting changes that occur in a small number of samples. We introduce the "outlier-sum" statistic, and compare it to both the *t*-statistic and COPA in a number of examples.

2. The outlier-sum statistic

Let x_{ij} be the expression values for genes $i = 1, 2, \ldots, m$ and samples $j = 1, 2, \ldots, n$. We assume that the samples fall into two groups. We think of group 1 as a normal or reference group, while group 2 is a disease group. Let C_k be the set of indices of the observations in group k, for $k = 1, 2$. The standard (unpaired) *t*-statistic is

$$T_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{s_i}. \tag{2.1}$$

Here $\bar{x}_{ik}$ is the mean of gene i in group k and s_i is the pooled within-group standard deviation of gene i.

As an alternative, we define the outlier-sum statistic as follows. Let med_i and mad_i be the median and median absolute deviation of the values for gene i. We first standardize each gene

$$x'_{ij} = (x_{ij} - \text{med}_i)/\text{mad}_{i1}. \tag{2.2}$$

This standardization puts all genes on the same scale to facilitate comparisons across genes.

Let $q_r(i)$ be the rth percentile of the x'_{ij} values for gene i, and $\text{IQR}(i) = q_{75}(i) - q_{25}(i)$, the interquartile range. Finally, note that values greater than the limit $q_{75}(i) + \text{IQR}(i)$ are defined to be outliers in the usual statistical sense.

The outlier-sum statistic is defined to be sum of the values in the disease group that are beyond this limit:

$$W_i = \sum_{j \in C_2} x'_{ij} \cdot I[x'_{ij} > q_{75}(i) + \text{IQR}(i)]. \tag{2.3}$$

Hence, W_i is large if there are many outliers in the disease group, or a few outliers with large values. If there are no outliers, then W_i is zero.

As an example, we generated 1000 genes and 30 samples, all values drawn independently from a standard normal distribution. Then, we added two units to gene 1 for four of the samples in the second group. We computed the *P*-value: the proportion of genes with score greater than that for gene 1, in absolute value. This process was repeated 50 times. The results for both the *t*-statistic and the outlier-sum statistic are shown in Figure 1. We see that the outlier-sum statistic yields smaller (more significant) *P*-values overall.

In real applications, one might expect negative as well as positive outliers. Hence, we define

$$W'_i = \sum_{j \in C_2} x'_{ij} \cdot I[x'_{ij} < q_{25}(i) - \text{IQR}(i)] \tag{2.4}$$

and set the outlier sum to the larger of $W(i)$ and $W'(i)$ in absolute value. We call this the "two-sided outlier-sum statistic," and illustrate its use in the skin data example of Section 4.

Note that the outlier-sum statistic is not symmetric in the classes. It explicitly looks for outliers in group 2, treating group 1 as a normal reference class. If finding outliers in group 1 is also of interest, then the procedure can be applied with groups 1 and 2 interchanged.

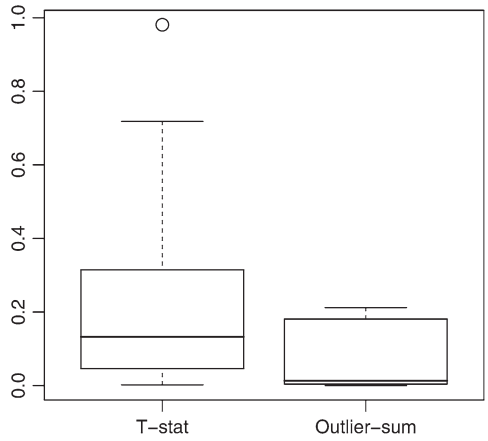

Fig. 1. Simulated example: *P*-values for gene 1 over 50 simulations.

Table 1. *Results of simulation study: median, mean, and standard deviation of P-values for gene* 1, *over* 50 *simulations*

	$k = 15$			$k = 8$			$k = 4$			$k = 2$		
	t	COPA	Outlier sum	*t*	COPA	Outlier sum	*t*	COPA	Outlier sum	*t*	COPA	Outlier sum
Median	0	0.293	0.110	0.012	0.100	0.105	0.119	0.094	0.098	0.250	0.182	0.100
Mean	0	0.332	0.106	0.029	0.151	0.094	0.164	0.156	0.093	0.303	0.274	0.100
SD	0	0.235	0.019	0.042	0.141	0.030	0.165	0.192	0.130	0.240	0.283	0.131

3. SIMULATION STUDY AND COMPARISON TO THE COPA STATISTIC

We carried out a small simulation study to assess the relative performance of the *t*-statistic, COPA, and the outlier sum. The COPA statistic (Tomlins *and others*, 2005) is defined as follows. All measurements for a gene are standardized by the overall median and median absolute deviation for that gene. Then the COPA statistic is the *r*th quantile of the data in the disease group. The authors use $r = 0.75, 0.90$, or 0.95. In our comparison below, we use the intermediate value of 0.90. In their paper, the authors apply the procedure to data from just the disease group. However, the standardization in the first step could also make use of a normal group, if available. We try both approaches in the simulation studies below.

We generated the data in the same way as in Figure 1. There are 1000 genes and 30 samples, all values drawn from a standard normal distribution. Then we added two units to gene 1 for k of the samples in the second group. We computed the *P*-value: the proportion of genes with score greater than that for gene 1, in absolute value (smaller values are better). This entire process was repeated 50 times. The values of k tried were 15, 8, 4, and 2. The median, mean, and standard deviation of the *P*-values are shown in Table 1.

When $k = 15$, so that all samples in group 2 are differentially expressed, the *t*-statistic performs the best. It continues to win when $k = 8$. But for smaller values of k, the outlier-sum statistic yields lower *P*-values, and has smaller standard deviation. The COPA statistic has consistently higher *P*-values than the outlier sum.

Table 2 shows the results when each method uses only the 15 samples from the disease class. Hence, the *t*-statistic in Table 2 refers to the one-sample *t*-statistic, and similarly for COPA and outlier sum.

Table 2. *Results of simulation study, one sample setting: median, mean, and standard deviation of P-values for gene* 1, *over* 50 *simulations*

	$k = 15$			$k = 8$			$k = 4$			$k = 2$		
	t	COPA	Outlier sum	*t*	COPA	Outlier sum	*t*	COPA	Outlier sum	*t*	COPA	Outlier sum
Median	0.000	0.565	0.125	0.022	0.558	0.125	0.124	0.290	0.123	0.279	0.442	0.124
Mean	0.001	0.553	0.169	0.052	0.548	0.140	0.212	0.344	0.113	0.346	0.420	0.110
SD	0.004	0.281	0.212	0.080	0.271	0.119	0.212	0.275	0.033	0.265	0.288	0.033

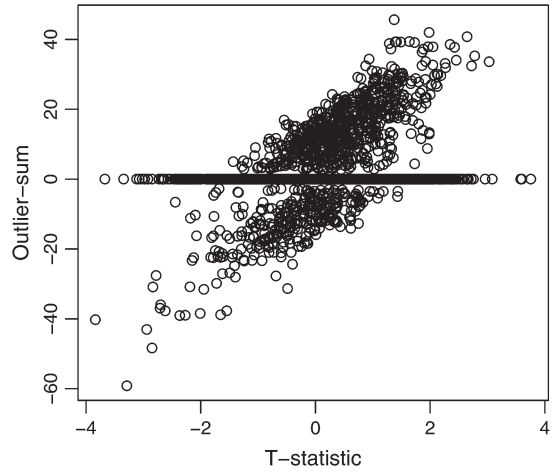

Fig. 2. Skin data: outlier-sum statistic versus the *t*-statistic. Note that many genes have outlier sums of zero.

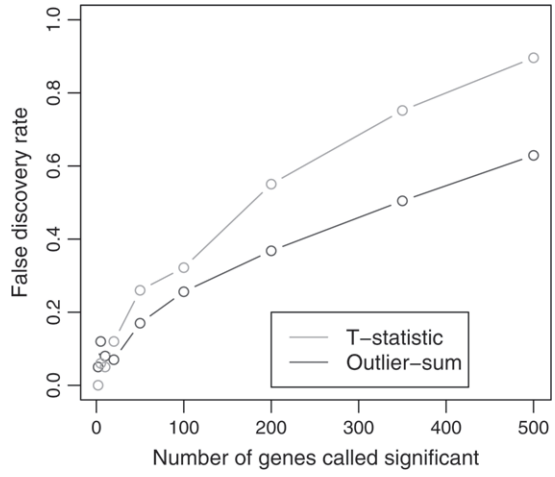

Fig. 3. Skin data: FDR versus the number of genes called significant, as the corresponding thresholds for each statistic is varied.

.

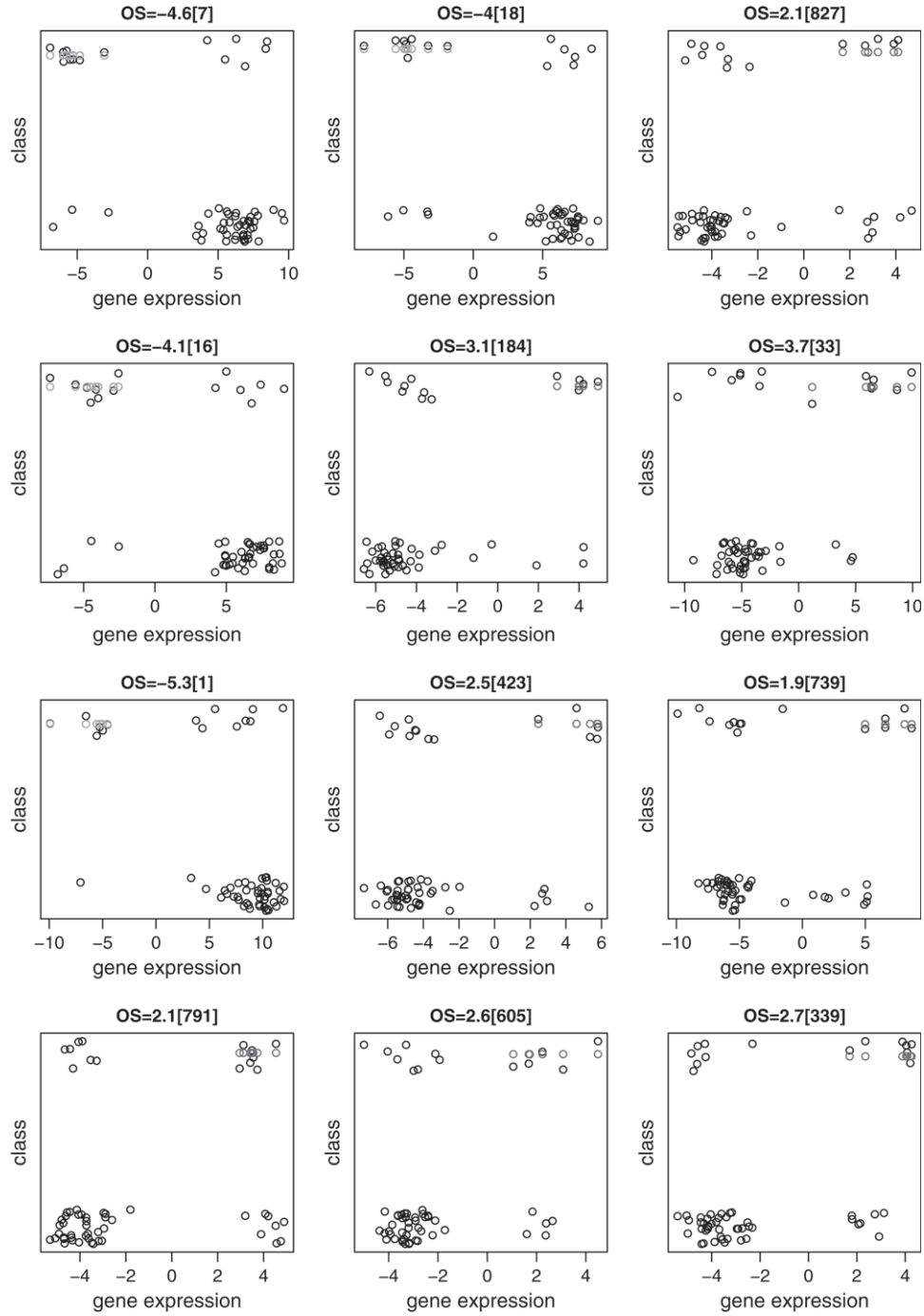

Fig. 4. Plots of the expression values in each class, for the 12 genes ranked highest by the outlier-sum statistic. The points have been "jittered" in the vertical direction, for clearer viewing. The number in brackets is the rank given to that gene by the *t*-statistic. The red points are identified as positive outliers; the green points are negative outliers.

The outlier sum performs a little worse than it did in Table 1, but still offers some improvement over the *t*-statistic when the number of outliers is small. The performance of the COPA statistic is noticeably worse in the one-sample case.

These experiments suggest that the outlier-sum statistic can provide a useful alternative to the *t*-statistic. With real data, one can estimate FDRs of both procedures to get an idea of which procedure is most informative for the data at hand. We illustrate this in Section 4.

4. Application to the skin data via the SAM method

In this example taken from Rieger *and others* (2004), there are 12 625 genes and 58 cancer patients: 14 with radiation sensitivity and 44 without radiation sensitivity. We applied the outlier-sum statistic within the SAM (Significance analysis of microarrays) approach (Tusher *and others*, 2001), using the group of 44 as the normal class. SAM estimates differential expression using the two-sample *t*-statistic and estimates FDRs via permutations of the class labels. Here we compare the *t*-statistic to the outlier-sum statistic in SAM. Since the data are from Affymetrix chips and have a wide range of expression, we first took cube roots. However, in practice a more careful preprocessing should be used, such as that provided by robust multi-chip analysis (Irizarry *and others*, 2003).

Figure 2 shows the outlier-sum statistic plotted against the *t*-statistic. These two scores are correlated but still differ substantially.

Figure 3 shows the FDR versus the number of genes called significant, as the corresponding thresholds for each statistic are varied. We see that the outlier-sum statistic has lower FDR near the right of plot, although the FDR may be too high there, for it to be useful in practice. Figure 4 shows the top 12 genes called by the outlier-sum statistic, plotted by group. The number in brackets is the rank given to that gene by the *t*-statistic, and we see that some of these genes are ranked quite low.

The outlier-sum statistic will appear in an upcoming version of the SAM package, available at http://www-stat.stanford.edu/∼tibs/SAM.

Acknowledgments

Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183. *Conflict of Interest:* None declared.

References

Allison, D., Cui, X., Page, G. and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55–65.

Dudoit, S., Yang, Y., Callow, M. and Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **97**, 111–39.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U. and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **2**, 249–64.

Lyons-Weiler, J., Patel, S., Becich, M. and Godfrey, T. (2004). Tests for finding complex patterns of differential expression in cancers: towards individualized medicine. *BMC Bioinformatics* **5**.

Rieger, K., Hong, W., Tusher, V., Tang, J., Tibshirani, R. and Chu, G. (2004). Toxicity from radiation therapy associated with abnormal transcriptional responses to DNA damage. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 6634–40.

TOMLINS, S. A., RHODES, D. R., PERNER, S., DHANASEKARAN, S. M., MEHRA, R., SUN, X.-W., VARAMBALLY, S., CAO, X., TCHINDA, J., KUEFER, R. *and others* (2005). Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *Science* **310**, 644–8.

TUSHER, V., TIBSHIRANI, R. AND CHU, G. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–21.

[*Received March 24, 2006; accepted for publication May 10, 2006*]