Improved detection of differential gene expression through the singular value decomposition

Robert Tibshirani *
Eric Bair †

Abstract

We propose a method for detecting differential gene expression that makes use of the singular value decomposition of the matrix of expression values. It looks for biological variation that correlates with the outcome variable, and when used in conjunction with the Significance Analysis of Microarrays (SAM) method can sometimes produce gene lists with lower false discovery rates.

1 Introduction

We consider methods for detecting differentially expressed genes in from a set of microarray experiments. Most existing methods use some measure of correlation measure between each gene and the outcome of interest. Consider for example the setting where we observe an expression profile and a possibly censored survival time for a set of patients. For each gene one can compute a score from Cox's proportional hazard's model, and then rank the genes according to this score. Genes whose absolute score is larger than some threshold are called "significant". Permutations of the survival times can be used to estimate the False Discovery Rate (FDR) of the resulting rule

*Dept. of Health Research & Policy, and Department of Statistics, Stanford University, Stanford, CA 94305. Email: tibs@stat.stanford.edu.

†Department of Statistics, Stanford University, Stanford, CA 94305. Email: ebair@stat.stanford.edu.

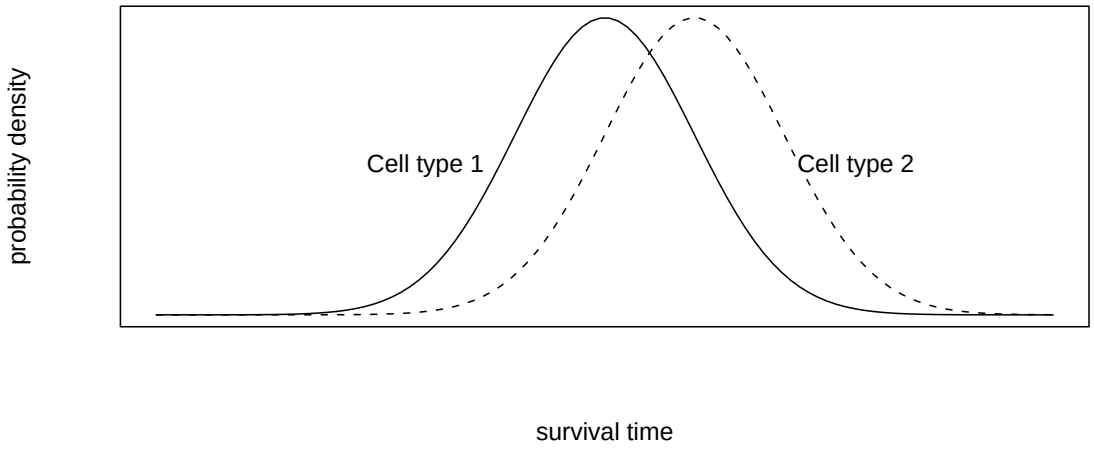

Figure 1: *Underlying conceptual model for Example 1.*

for each threshold value and the FDR can help determine the best choice of threshold.

A number of authors have proposed methods for detecting differential gene expression, including Dudoit et al. (2000), Newton et al. (2001) and Kerr et al. (2000). The preceding recipe describes the Significance of Microarrays (SAM) procedure (Tusher et al. 2001). In this short paper, we describe how eigengenes and eigenarrays can be used to improve the detection of differential gene expression.

As a motivating example and proof of concept, we generated data on 1000 genes and 40 samples. All expression values were generated as standard Gaussian, except for genes 1–50 in samples 21–40 which have mean 2.0. We think of samples 1–20 and 21–40 as representing two different cell types, with cell type 2 characterized by higher expression in the first 50 genes. An uncensored survival time was also generated for each sample, with a mean 1.0 units higher in samples 21–40 than in samples 1–20. Thus patients with cell type 2 tend to live longer than those with cell type 1, but there is considerable overlap in the two sets of survival times. We would like to detect that the first 50 genes are differentially expressed in these data. Figure 1 shows the underlying conceptual model.

Since the survival times differ by only 1 unit across the samples, methods like SAM that correlate expression with outcome directly will perform only moderately well in this example. We can do better through use of the singular value decomposition of the matrix of expression values. This decomposition produces a set of "eigengenes"- each being a linear combination of genes, showing the largest variation across the set of samples[1]. If there are n samples, there are n eigengenes, ordered from largest to smallest variance. Corresponding to the eigengenes are the "eigenarrays", each one being a linear combination of the expression profiles of the samples (arrays). The eigenarrays are ordered from largest to smallest variance across the genes.

The top panels of Figure 2 show the first eigengene and eigenarray for this example. The eigengene picks up the variation due to cell type, while the eigenarray shows that the first 50 genes are very different from the rest. The Cox scores for each gene are shown in the bottom left panel. The scores for the first 50 genes seems to be a little lower on average, but they do not clearly stand out as being different.

Our proposal in this paper is to find the eigenarray most correlated with the set of scores, and then do a least squares fit of the scores on the eigenarray to obtain a new improved set of scores. In this example, the most correlated eigengene is the first one (shown in the top left) and the least squares fit of the scores in shown in the bottom right. The scores for the first genes are now clearly different from the rest.

Here our idea clearly helped. But this will not always be the case. The method seems to require that there be strong biological variation correlated with survival time, and detectable from gene expression. Hence we need some objective way to determine whether the use of the eigenarray helps in a given example. Fortunately, estimation of false discovery rates via permutations of the survival times (as done in SAM), can be carried out in exactly the same way for the new procedure. This FDR can be used to help determine whether use of the eigenarray is helpful for a given problem.

The left panel of Figure 3 shows the actual number of falsely called genes as a function of the number of called genes, as the threshold is varied. Results are shown both from SAM and the proposed procedure (in the examples we label the new procedure as "eigenSAM"). The FDR is the slope of the curves in this plot. Computation of these quantities uses the knowledge of the

[1]The terms "eigengenes"and "eigenarrays" were coined by Alter et al. (2000). They are the singular vectors, or principal components of the expression matrix

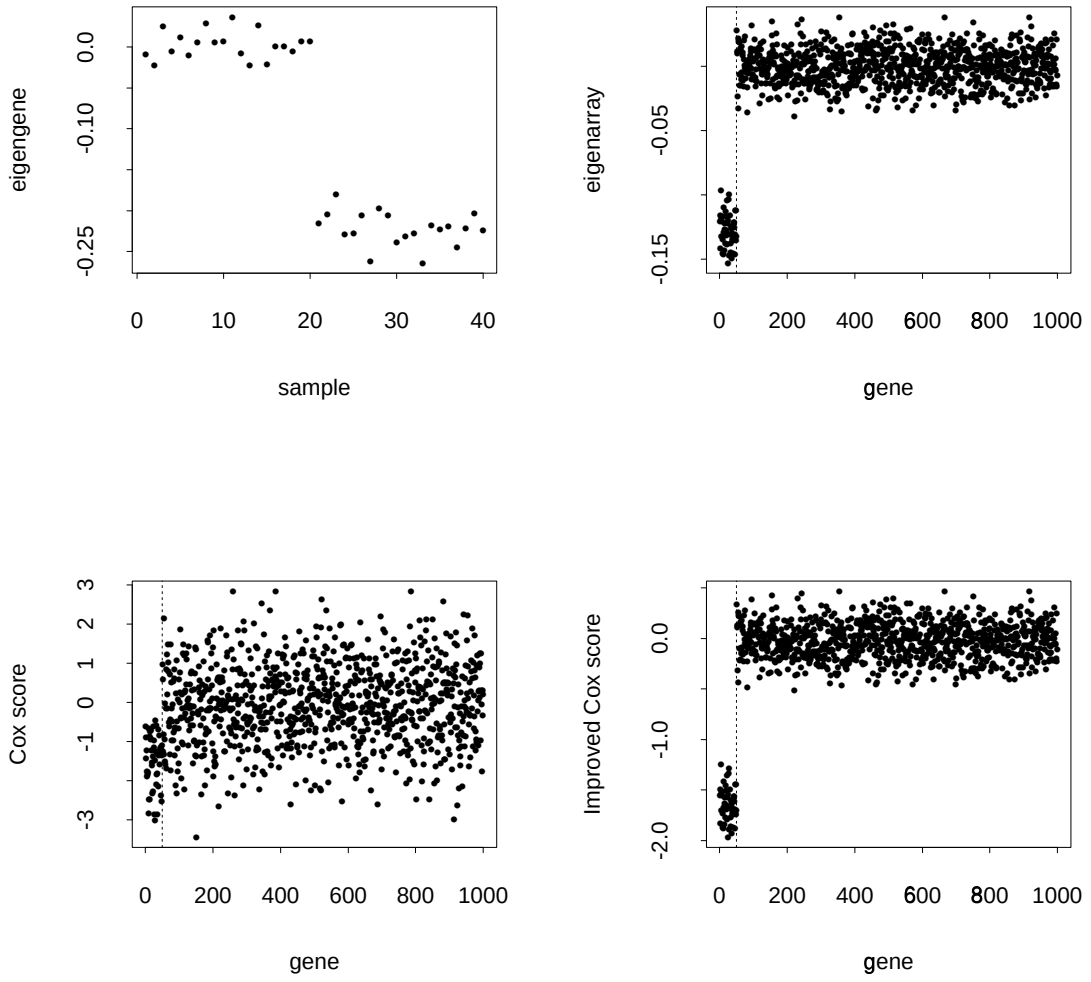

Figure 2: *Results for toy example. Top row shows the leading eigengene and eigenarray. Bottom left panel shows the Cox scores for each gene. Bottom right panel shows the improved scores, produced by a least squares fit of the scores onto the eigenarray in the top right panel. The vertical lines mark the boundary between the first 50 (non-null) genes and the rest.*

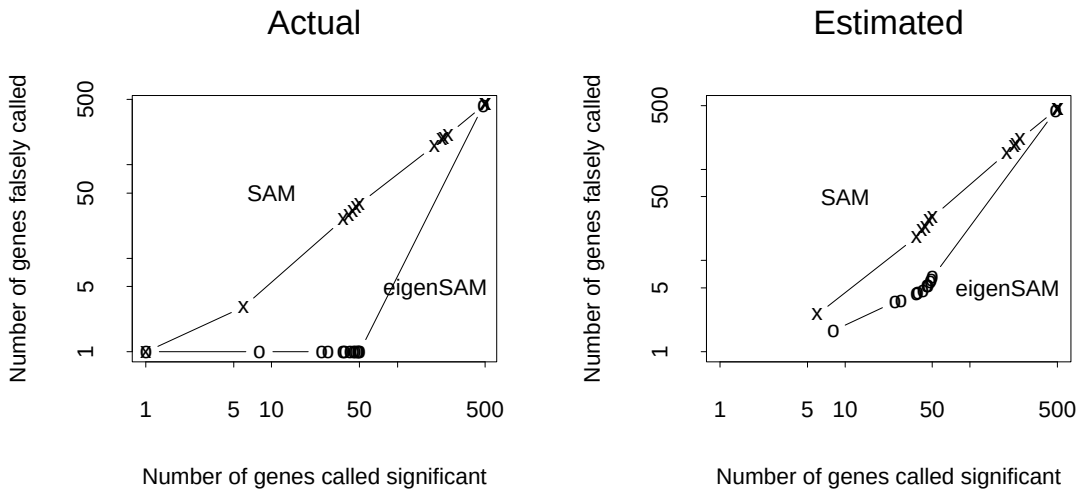

Figure 3: *Results for toy example*

first 50 genes are truly differentially expressed, and the others are not. The top right panel shows the estimated version of these quantities, using data permutations. In both cases the FDR is seen to be considerably lower for the proposed procedure.

Can the use of eigenarrays give misleading results? We next tried to "break" the procedure. We generated data on $i = 1, 2, \ldots 1000$ genes and $j = 1, 2, \ldots 40$ samples, where $y_j \sim N(10, 1)$ for $j \leq 20$, $y_j \sim N(12, 1)$ for $j > 20$, and $x_{ij} \sim N(\mu_{ij}, 1)$ where

$$\begin{aligned} \mu_{ij} &= \beta_j + \tau_{ij} \\ \beta_j &= 2 \text{ for } j \text{ odd, and zero otherwise} \\ \tau_{ij} &= 1 \text{ for } i \leq 50 \text{ and } j > 20, \text{ and 0 otherwise} \end{aligned} \tag{1}$$

Hence the survival time is higher in the second group of 20 patients, as is the expression of the first 50 genes. But for all genes there is a much stronger variation in expression values that is uncorrelated with the survival difference.

Figure 4 shows the actual and estimated number of falsely called genes. As we see, the use of the eigenarray has improved both the actual and estimated false discovery rates. The procedure has in effect removed the extraneous variation (similar to blocking in an analysis of variance) and this results in improved accuracy.

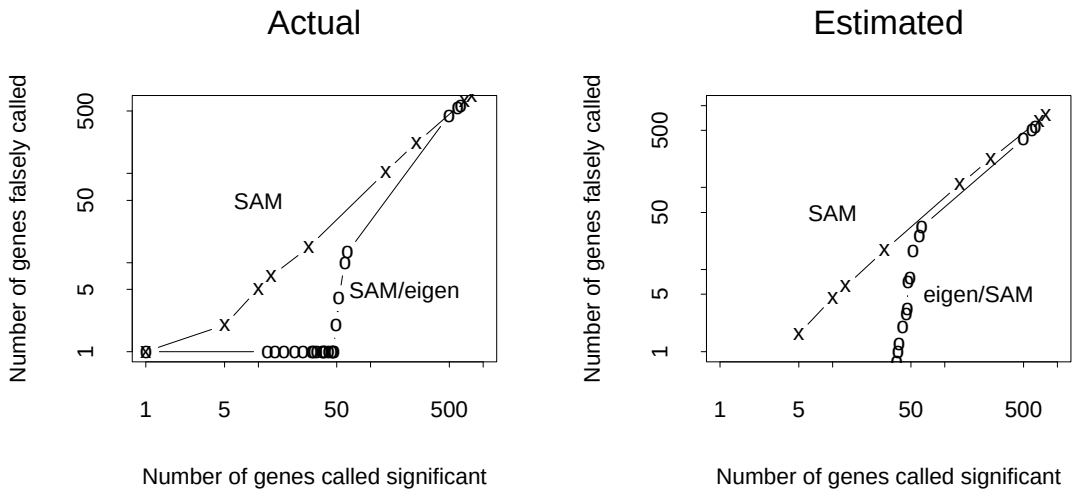

Figure 4: *Results for second toy example*

In other experiments we were unable to concoct an example where use of the eigenarray increased the FDR. In some cases it didn't help, but this also was clear from the estimated false discovery rate. In practice, one can simply try the procedure using the false discovery rate as a guide.

2 Details

Let X be the $n \times p$ matrix of expression values, for p genes and n samples. Assume the columns of X are centered. Denote the singular value decomposition of X by

$$X = UDV^T \tag{2}$$

where U is an $p \times n$ orthonormal matrix with columns $u_1, u_2 \ldots u_n$, V is an $n \times n$ orthonormal matrix with columns $v_1, v_2, \ldots v_n$ and D is an $n \times n$ diagonal matrix with diagonal elements equal to the singular values θ_j of X. We assume $p \leq n$. Thus

$$Xv_j = \theta_j u_j;\ j = 1, 2, \ldots n \tag{3}$$

where u_j and v_j are vectors of length p and n respectively. Alter et al. (2000) call u_j the "eigenarrays" , and v_j the "eigengenes" of X. We describe our idea in the case of a survival outcome and associated Cox scores, but the same idea can be applied to other outcome types such as two-class, paired, quantitative etc. Here are the steps:

1. Compute the eigenarrays u_j and eigengenes v_j of X, $j = 1, 2, \ldots n$.
2. Compute the Cox scores $d = (d_1, d_2, \ldots d_p)$ for each of the p genes.
3. Find the eigenarray with largest absolute correlation with d:

$$k = \text{argmax}_i |\text{corr}(d, u_i)|$$

4. Compute new scores d' equal to the least squares fit of d on u_k. Since each u_i is a unit vector with mean zero, this has the simple form

$$d' = \bar{d} + \langle d, u_k \rangle \cdot u_k$$

We then use the new scores d' in the same way that the original scores d were used: we rank the genes by their scores to determine their relative significance.

Suppose that instead of a survival time outcome, we have a quantitative outcome y, and the gene score is chosen to be simply the inner product of each row of the expression matrix with y. Then the above procedure is equivalent to the following:

- Find the eigengene v_k most correlated with y
- Replace y by its least squares fit $\hat{y}$ on v_k.
- Compute new scores d' as the inner product of each gene with $\hat{y}$.

Now one doesn't typically use the simple inner product to score a gene. One would instead use a standardized inner product, i.e. the least squares slope divided by its estimated standard error. However the point of this is to show that replacing the scores by their least squares fit on the most correlated *eigenarray*, is approximately equivalent to replacing the outcome variable with its least squares fit on the most correlated *eigengene*.

Estimation of the false discovery rate is done via permutations of the survival times, as in SAM. It uses the eigengenes and eigenvalues of X from above: these do not have to be recomputed. Here are the steps:

1. Randomly permute the survival times.
2. Compute the Cox scores $d^* = (d_1^*, d_2^*, \ldots d_p^*)$ for each of the p genes.
3. Find the eigenarray with largest absolute correlation with d:

$$k* = \text{argmax}_i |\text{corr}(d^*, u_i)|$$

4. Compute new scores $d^{*'}$ equal to the least squares fit of d^* on u_{k*}.
5. Repeat steps 1-4 many times, to get the expected quantiles of the null scores d^*.

The resulting quantiles are used to estimate the cutpoints and the corresponding FDR, as detailed in Tusher et al. (2001).

3 Lymphoma example

We applied this idea to a dataset on diffuse large cell lymphoma from Rosenwald et al. (2002). There are 7399 genes and 240 patients in the training set, and 80 patients in the test set. Time until death (possibly censored) is available for each patient. Figure 5 shows the estimated performance of SAM and eigenSAM on the training set. Use of the eigenarray seems to have improved the FDR considerably.

To assess performance on the test set, we computed Cox scores for the test set and declared any gene with score larger than 2.0 in absolute value to be a truly significant (non-null) gene. We then varied the threshold for SAM and eigenSAM in the training set, and computed the number of truly significant genes each time. The results are shown in Figure 6. The new procedure find more significant genes in the test set, for the same number of genes called significant in the test set, although neither method does very well overall. For example with 200 genes called, SAM finds 9 truly significant genes and eigenSAM finds 20. These are listed in Table 1 (note that there are some duplicates- the same gene with different clones). Only one gene is common to the two lists, and eigenSAM has found more genes and genes with lower p-values.

The eigenarray most correlated with the Cox gene scores was the one with third highest variance u_3, with a correlation of .68. Figure 7 shows a plot

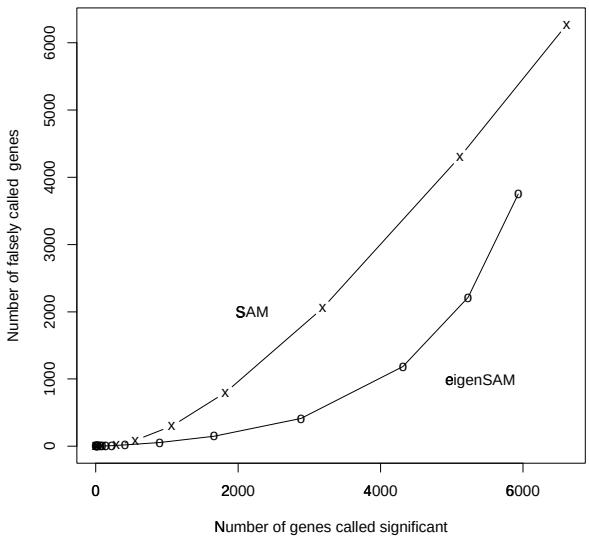

Figure 5: *Training set results for lymphoma example.*

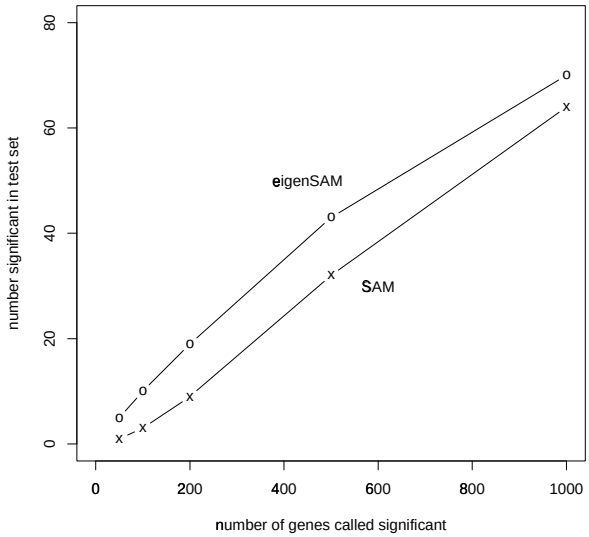

Figure 6: *Test set results for lymphoma example.*

Table 1: *Genes significant in test set, from SAM and eigenSAM*

Test set p-value	Gene
	SAM
0.0025	—U10485—*AA457051—Hs.40202—lymphoid-restricted membrane protein
0.0048	—J03040—*W46959—Hs.111779—secreted protein, acidic
0.0065	——*AA828425—Hs.291892—EST
0.0079	—M25393—*AA193262—Hs.82829—protein tyrosine phosphatase
0.0158	—M61906—*N69643—Hs.6241—phosphoinositide-3-kinase, regulatory subunit
0.0192	————LC33743
0.0198	————LC32442
0.0203	—X52142—*W44416—Hs.251871—CTP synthase
0.0211	—V00568——Hs.79070—v-myc myelocytomatosis viral oncogene homolog (avian)
	eigenSAM
0.0002	—X61118—*AA280651—Hs.184585—LIM domain only 2 (rhombotin-like 1)
0.0003	—U11732—*AA831368—Hs.169081—ets variant gene 6 (TEL oncogene)
0.0006	—AF178632—*AA827145—Hs.6048—fem-1 homolog b (C. elegans)
0.0007	—X61118—*AA280651—Hs.184585—LIM domain only 2 (rhombotin-like 1)
0.0017	—X61118—*AA261902—Hs.184585—LIM domain only 2 (rhombotin-like 1)
0.0025	—U10485—*AA457051—Hs.40202—lymphoid-restricted membrane protein
0.0032	——*AA832051—Hs.369936—ESTs
0.0063	——*AA731512——
0.0071	————LC19314
0.0077	—M15395—*AA287298—Hs.83968—integrin, beta 2 (antigen CD18 (p95),
0.0080	—M26004—*AA262317—Hs.73792—complement component (3d/Epstein Barr virus)
0.0080	—M14745—*W63749—Hs.79241—B-cell CLL/lymphoma 2
0.0087	—U07620—*R39221—Hs.151051—mitogen-activated protein
0.0108	—D89289— AA056991—Hs.118722—fucosyltransferase 8 (alpha (1,6)
0.0111	—X12654—*AA291398—Hs.84746—chromosome condensation 1
0.0131	—X55188—*R76698—Hs.56729—lymphocyte-specific protein 1
0.0134	—U44403—*H29540—Hs.75367—Src-like-adaptor
0.0144	—M26004—*AA465705—Hs.73792—complement component
0.0195	—BC007655—*R72567—Hs.267819—protein phosphatase 1

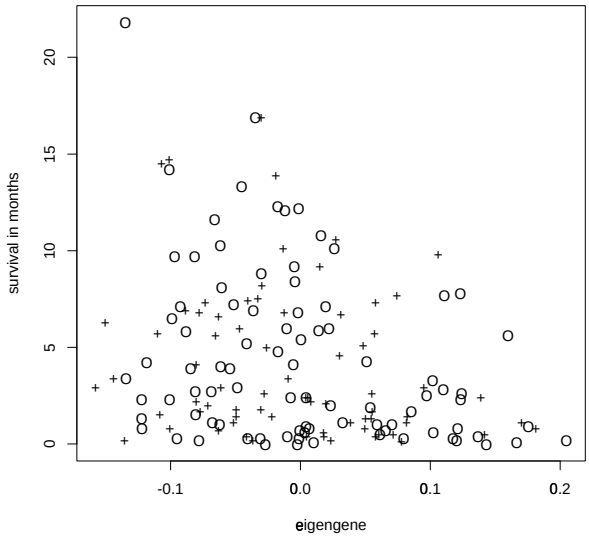

Figure 7: *Survival time versus the eigengene most correlated with survival time. Censored observations are indicated by a "+".*

of survival time versus the corresponding eigengene v_3. Larger values of the eigengene correlate with poorer survival. Examination of the characteristics of the samples corresponding to low and high values of this eigengene, might reveal biological insight as to underlying mechanisms of the disease.

4 Discussion

The proposal of this paper can be applied to outcome measures other than survival times, including categorical, quantitative or paired samples. This proposal will be offered as an option in a future version of the SAM package available at `http://www-stat.stanford.edu/~tibs/SAM`

The same general idea can potentially be applied to a different problem: sample classification from gene expression profiles. This is a topic for future research.

References

Alter, O., Brown, P. & Botstein, D. (2000), 'Singular value decomposition for genome-wide expression data processing and modeling', *Proceedings of the National Academy of Sciences* **97**(18), 10101–10106.

Dudoit, S., Yang, Y., Callow, M. & Speed, T. (2000), Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Unpublished, available at http://www.stat.berkeley.edu/users/sandrine.

Kerr, M., Martin, G. & Churchill, G. (2000), 'Analysis of variance for gene expression microarray data', *Journal of Computational Biology* **7**, 819–837.

Newton, M., Kendziorski, C., Richmond, C., Blatter, F. & Tsui, K. (2001), 'On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data', *Journal of Computational Biology* **8**, 37–52.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lync, h. J. C., Vose, J., Armitage, J. O., Montserrat, E., Lopez-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., H., H., Krajci, P., Stokke, T. & Staudt, L. (2002), 'Use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma', *N. Engl J Med.* pp. 1937–1947.

Tusher, V., Tibshirani, R. & Chu, G. (2001), 'Significance analysis of microarrays applied to transcriptional responses to ionizing radiation', *Proc. Natl. Acad. Sci. USA.* **98**, 5116–5121.