

Data-adaptive test statistics for microarray data

Sach Mukherjee¹, Stephen J. Roberts², Mark van der Laan³.

Keywords: differential expression, microarrays, machine learning, multiple testing

1 Introduction.

In recent years, a great deal of research has focused on the statistical task of selecting differentially expressed genes from microarray data ('gene selection'). In this poster, we present a novel gene selection procedure in which test statistics are learned from data, using a simple notion of reproducibility in selection results as the learning criterion. Reproducibility, as we define it, can be computed without any knowledge of the 'ground-truth', but nonetheless provides an asymptotically valid guide to expected loss under the true data generating distribution. Empirical results on simulated and real microarray data demonstrate the accuracy and robustness of our method compared with two widely-used test statistics.

2 Theory.

Our aim is to learn test statistics for gene selection from data. Ideally, we would like to choose a test statistic which minimises expected loss under the true data-generating distribution. However, in gene selection, we can neither adequately characterise the underlying distribution, nor compute error on given data (since we do not know which genes are truly differentially expressed), so we cannot hope to directly minimise expected loss. Instead we use as our learning criterion a simple notion of *reproducibility*. We define the reproducibility of a test statistic as the number of genes in common between gene-lists obtained using that statistic from a pair of datasets drawn from the same underlying distribution, and estimate it using a bootstrap procedure. Formal results show that asymptotically, reproducibility is anti-correlated with expected loss (under certain quite benign conditions), such that maximising reproducibility is equivalent to minimising expected loss. Thus, if $R(f, D)$ is the reproducibility of test statistic f on data D , we will use the function f^* as our test statistic:

$$f^* = \arg \max_{f \in \mathcal{F}} R(f, D) \quad (1)$$

Where \mathcal{F} is a suitably chosen family of test statistics. For the problem of selecting differentially expressed genes from two-class expression data we use the family of statistics \mathcal{F} defined by $(d_i + k_1)/(k_2 \times \hat{\sigma}_i + k_3)$. Here, d_i refers to the absolute difference in sample means between the two classes for gene i , and $\hat{\sigma}_i$ to the standard deviation for gene i . The k s are parameters to be learned and are constrained in the following way: $k_1, k_3 \in [0, 5]$; $k_2 \in \{0, 1\}$.

3 Results.

Simulated data were generated (following a similar procedure to [2]) under a series of challenging conditions, including differing variances, non-normality and bi-modality. In each case, we generated 200 datasets, each with 1025 "genes", of which 25 were truly differentially expressed. We used these datasets to obtain ROC curves and box-plots of the number

¹Department of Engineering Science, University of Oxford, U.K. E-mail: sach@robots.ox.ac.uk

²Department of Engineering Science, University of Oxford, U.K. E-mail: sjrob@robots.ox.ac.uk

³Division of Biostatistics, U.C. Berkeley, U.S.A. E-mail: laan@stat.berkeley.edu

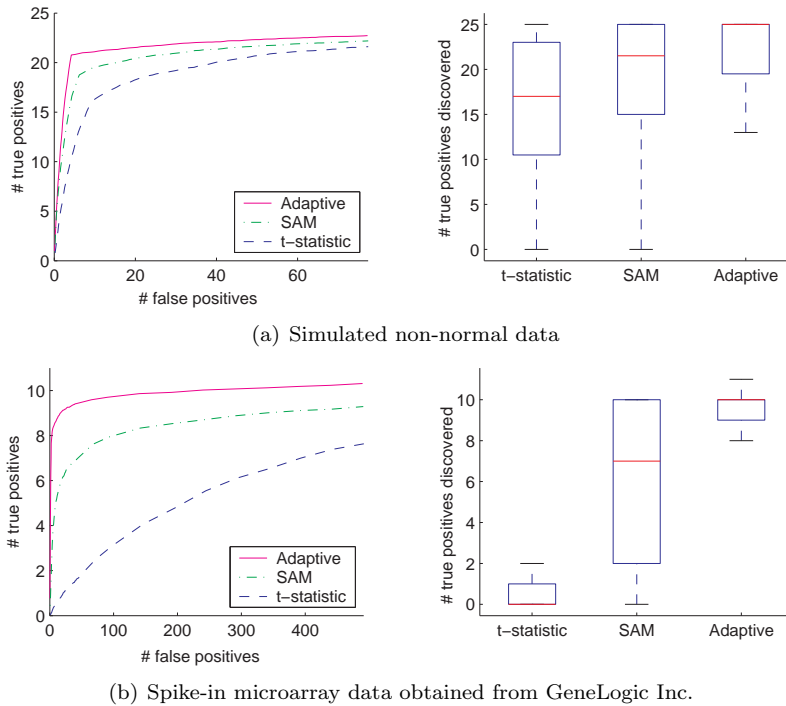


Figure 1: Results on simulated and benchmark microarray data. The box-plots show the number of true positives discovered when the number of genes selected is fixed.

of true positives discovered (when the top 25 genes are selected), for our method as well as the t-statistic and SAM [1]. In most cases, we found our method outperformed both the t-statistic and SAM, with the improvement being very significant in many cases. Interestingly, under some conditions (equal variances, Normal data, moderate sample-sizes), our method ‘learned’ the t-statistic; but, when the data-generating distributions departed significantly from the canonical model, our method continued to perform well, usually learning a close-to-optimal member of the family \mathcal{F} defined above. Figure 1(a) shows the results of one of our experiments; here, the underlying model is significantly non-normal.

We also applied our method to a spike-in study conducted by GeneLogic (Gaithersburg, MD); the ROC curves and box-plots shown in Figure 1(b) were obtained using a resampling procedure. Again, our method clearly outperforms both the t-statistic and SAM.

In conclusion, we have proposed a data-adaptive procedure for selecting differentially expressed genes from microarray data. The method is both principled and effective, and outperforms widely-used methods on simulated as well as real microarray data.

References

- [1] Tusher, V. Tibshirani, R. and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA*, 98:5116–5121.
- [2] Lonnstedt, I. and Speed, T. P. 2002. Replicated microarray data. *Statistica Sinica*, 12:31–46.