AUSTRIAN JOURNAL OF STATISTICS
Volume 32 (2003), Number 3, 225-238

Cost-effective Screening for Differentially Expressed Genes in Microarray Experiments Based on Normal Mixtures

Jörg Rahnenführer [1]
Max-Planck Institut für Informatik, Saarbrücken, Germany

Andreas Futschik [1]
Department of Statistics, University of Vienna, Austria,

Abstract: Microarray experiments allow the monitoring of expression levels for thousands of genes simultaneously. Based on data obtained from the co-hybridization of two mRNA samples, a frequent goal is to find out which genes are differentially expressed. For this purpose, we propose to estimate the distribution of popular test statistics by a mixture of normal distributions. These statistics are calculated for each gene separately. A Bayes classifier is then used to decide upon differential expression. The cut-off for the classifier is chosen according to the number of false positives and negatives when applied to realistic data generating models. In particular, we generate data from a mixture model and from an Empirical Bayes model. By comparing the numbers of false decisions for various test statistics in the context of the considered models, we investigate which of the statistics are particularly suitable with our approach.

Zusammenfassung: Mit Microarray-Experimenten wird die Expression von Tausenden von Genen gleichzeitig gemessen. Basierend auf Daten von einer gemeinsamen Hybridisierung von zwei mRNA-Proben ist es häufig das Ziel, die differentiell exprimierten Gene zu identifizieren. Wir schlagen vor, die Verteilung von beliebten Teststatistiken durch eine Mischung von Normalverteilungen zu schätzen. Die Statistiken werden dabei für alle Gene einzeln berechnet. Zur Bestimmung der differentiell exprimierten Gene wird ein Bayes-Klassifikator verwendet. Dessen cut-off wird anhand der Anzahl von falsch Positiven und falsch Negativen bei Anwendung auf realistische datenerzeugende Modelle gewählt. Insbesondere generieren wir Daten von einem Mischungsmodell und von einem Empirical-Bayes-Modell. Durch den Vergleich von Anzahlen von falschen Entscheidungen für verschiedene Teststatistiken im Kontext der vorgeschlagenen Modelle wird untersucht, welche Statistiken für unseren Ansatz am Besten geeignet sind.

Keywords: cDNA Microarray, Differential Expression, Mixture Model, Bayes Classifier, Comparison of Test Statistics.

1 Introduction

In cDNA microarray experiments, the gene expression of several thousand genes is measured simultaneously. DNA sequences of genes of interest are printed on a glass micro-

[1] Parts of this work were done at the Department of Statistics, University of California at Berkeley, USA.

scope slide using a robotic arrayer. Then the abundance of these DNA sequences in two mRNA samples is compared. The two mRNA samples are labelled with two different dyes, typically one with green-fluorescent Cy3 dye and the other with red-fluorescent Cy5 dye. They are mixed and co-hybridized on the microarray slide. The signals are read using a fluorescent imaging system like a con-focal scanner. In many microarray experiments the key question is which genes are differentially expressed in the two original samples. The underlying paradigm is that a significant shift in gene expression also has a biological meaning.

The identification of differentially expressed genes is usually performed in three steps. First, image analysis techniques are used to localize the probes on the microarray and to estimate the expression levels for each spot, see for instance Yang et al. (2002) or Bozinov and Rahnenführer (2002). Next, the obtained data are normalized to remove systematic variation (Yang et al., 2001). Finally, a decision has to be made which genes are likely to be differentially expressed. A simple ad hoc way to screen for such genes is to identify all those with more than a c-fold shift in the log expression ratio, where c is some fixed constant. Alternatively one can declare the top k genes (for instance $k = 100$) with respect to some test statistic as differentially expressed. However, the choice of such a constant or number is highly dependent on the specific experiment, the underlying biology and the expected number of differentially expressed genes. Thus a specific predetermination of this number is difficult to justify in general.

Therefore, alternative approaches to identify differentially expressed genes have been considered in the literature. Newton et al. (2001) propose a parametric model for single-slide data with only one observation pair per gene. Dudoit et al. (2000) consider the use of resampling based multiple tests, and Lönnstedt and Speed (2002) investigate an Empirical Bayes model. In a case study, Efron et al. (2001) propose a nonparametric mixture model for the identification of differentially expressed genes.

In this context, it would also be of interest to get some idea, how many differentially expressed genes are missed and how many genes are incorrectly declared to be differentially expressed. To address the second issue, Tusher et al. (2001) propose an estimate of the false discovery rate that is connected to their identification method relying on modified t-tests. Pan (2002) presents a comparative review of statistical methods for discovering differentially expressed genes. He compares three methods that are all based on a two sample t-test statistic or minor variations, they differ mainly in the way the significance level is associated with the corresponding statistic. Other methods from Efron et al. (2001) and Tusher et al. (2001) are briefly mentioned in this paper. Both estimate the null distribution from the data. Pan's comparison is done by an evaluation on one real data set with unknown true parameters.

In contrast to this study, we propose to fit a normal mixture model to popular test statistics calculated from microarray data. Based on the fitted mixture components, a Bayes classifier can be used to decide for differential expression. It turns out that three mixture components lead already to reasonable performance in many cases. Since the cut-off point of the classifier may be hard to choose in practice, the number of false discoveries and non-discoveries is provided for simulated data from realistic models, where it is known in advance, which genes are differentially expressed. We consider in particular a mixture model described in the next chapter and an Empirical Bayes model of Lönnstedt

and Speed (2002). Then numbers of false discoveries and non-discoveries are also compared for different popular statistics, leading to recommendations for test statistics in our context.

2 Proposed Approach

We focus on the case where n repeated measurements are available for p genes. For gene i and replicate j, we have two expression intensities $R_{i,j}$ and $G_{i,j}$ for the probes to be compared. Usually the decision concerning differential expression of gene i is based on some statistic T_i calculated from the vector of log-ratios $M_i = \log_2 \frac{R_{i,j}}{G_{i,j}}$ with $1 \leq i \leq p$ and $1 \leq j \leq n$.

Our approach is to consider a finite mixture model for the statistics T_i. We assume that T_i is generated according to one of $k+1$ probability distributions $P_0, \ldots, P_k$ with cumulative distribution functions $F_0, \ldots, F_k$ and densities $f_0, \ldots, f_k$. The unknown probability that T_i is distributed according to P_l is denoted by π_l for $l = 0, \ldots, k$.

Genes that are not differentially expressed are assumed to lead to statistics T_i generated from P_0 with $E_{P_0} T_i = 0$. For differentially expressed genes, let T_i be distributed according to one of the distributions $P_1, \ldots, P_k$ having nonzero expected values. In our simulations, we will focus on the situation where there are only two alternative distribution, P_1 with $E_{P_1} T_i < 0$ and P_2, where $E_{P_2} T_i > 0$. Since the difference in expression will, in general, vary across genes, this is a simplification of reality which may not always be adequate. In such situations, the model fit can always be improved by adding further mixture components, and methods for choosing an appropriate number k of components exist in the literature (McLachlan and Peel, 2000). However, for the only realistic data generating model (Lönnstedt and Speed, 2002) proposed so far to our knowledge, our simulations indicate that three mixture components perform already quite well.

Incorrect decisions usually lead to costs, denoted by λ_0 for a false discovery of a non-differentially expressed gene, and by $\lambda_1, \ldots, \lambda_k$ for an undetected differentially expressed gene with true distribution $P_1, \ldots, P_k$.

Figure 1 shows the decomposition of a mixture of three normal densities. The dashed line represents the density for non-differentially expressed genes with mean 0, the dash-dotted lines stand for two groups of up-regulated and down-regulated genes, respectively. The sum (solid line) has longer tails than a single normal distribution, which is known to be typical for microarray data.

If the components of the mixture model $\pi_0, \ldots, \pi_k$ and $P_0, \ldots, P_k$ were completely known, it would be easy to develop a cost-optimal classification procedure among all rules based on T_i. Indeed, such a procedure is given by the Bayes classifier

$$\varphi(\cdot) = \frac{\lambda_1 p_1 f_1(\cdot) + \lambda_2 p_2 f_2(\cdot) + \cdots + \lambda_k p_k f_k(\cdot)}{\lambda_0 p_0 f_0(\cdot)},$$

where a gene is identified as differentially expressed when $\varphi > 1$. For details see for instance Devroye et al. (1996).

In practice the densities are unknown and have to be estimated. For this purpose, Gaussian mixture models seem to be a good and flexible candidate, since any continuous

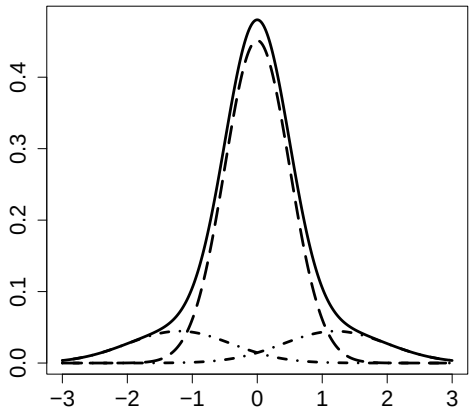

Figure 1: Densities of three normal distributions and of their mixture

distribution can be approximated arbitrarily well by a Gaussian mixture. This concept is utilized for instance by kernel density estimates based on Gaussian kernels. Besides the component densities, the costs $\lambda_0, \cdots, \lambda_k$ also determine the decision of the Bayes classifier. For the three component mixture situation with $\lambda_1 = \lambda_2$, Figures 2-5 provide guidelines concerning their choice. Given a realistic scenario model, the constants can be determined by choosing an acceptable trade-off between false positives and false negative decisions.

In our simulations, we used two different models to compare the quality of test statistics for microarray data. The first is the model proposed by Lönnstedt and Speed (2002), where replicates for gene i are generated from a normal $N(\mu_i, \sigma_i^2)$ distribution. The variances σ_i^2 are drawn from an inverse gamma prior distribution. By this we mean that d/σ_i^2 has a $\Gamma(\nu, 1)$ distribution with density

$$\frac{1}{\Gamma(\nu)} x^{\nu-1} e^{-x}.$$

(See Section 3 for the actual choice of the parameters ν and $d = na/2$.) The means μ_i are either set equal to zero or drawn from a normal prior distribution $N(0, c\sigma_i^2)$ with a fixed constant c. The second model analyzed is the normal mixture model described above.

Both models lead to log ratios that are independent for different genes. The main reason for this assumption is that in practical experiments nothing is known concerning the dependence structure. It should be noted that the results of Section 3 concerning expected numbers of false decisions are not affected by dependencies. The spread as displayed in our box plots should be interpreted with caution however, if strong dependencies are considered possible.

3 Simulations

In our simulations we consider microarrays containing $k = 6000$ genes. We assume that $n = 8$ replicate measurements are available for each gene. We consider both a small ($k_1 = 60$) and a large ($k_1 = 600$) number of differentially expressed genes. We carried out 100 simulations for every algorithm and parameter constellation. The data (i.e. the log-ratio vectors $M_i = (M_{ij})_{j=1}^n$) were generated according to the two models discussed in the last paragraph, namely the model of Lönnstedt and Speed (2002) and the normal mixture model.

For both models, gene variances σ_i^2 were drawn from an inverse gamma distribution with parameters $\nu = 2.8$ and $a = 0.04$, the constant c for differentially expressed genes in the Bayes model was set to $c = 1.2^{-1}$. This parameter choice is justifiable, since all values were obtained as estimates for a real data set analyzed by Lönnstedt and Speed. The parameters explicitly specify the original model introduced in this paper. Although the parameter values $\nu = 2.8$ and $a = 0.04$ are realistic, it is important to check, if the results hold across other choices. As an example of an unfavorable situation with significantly higher gene variances we chose $\nu = 2.5$ and $a = 0.06$. This leads to distributions usually not observed in real microarray experiments. Figure 2 shows the densities for standard deviations for both parameter sets, both on the original and on a log-scale.

In the mixture model with normal components, we assumed an individual normal distribution for the log ratios M_{ij} for each gene i. The means were chosen from one of the three values $\mu = -0.2,\ 0,\ 0.2$ according to fixed mixture probabilities. These probabilities were either set to $0.05, 0.9, 0.05$ or to $0.005, 0.99, 0.005$, representing 10% or 1% differentially expressed genes, respectively. Both models produce realistic so called MA-plots. In such plots the log ratios of single genes are plotted against their average intensities. The plots are popular, intuitive and the basis of various normalization methods.

We compared four different test statistics:

1. Mean log-ratios $\bar{M}_i = (\sum_{j=1}^n M_{ij})/n$.

2. T–test statistics $t_i = \bar{M}_i / SD(\bar{M}_i)$, where $SD(\bar{M}_i)$ denotes the standard deviation of the replicates $(M_{ij})_{j=1}^n$.

3. The modified t–statistics s_i proposed by Tusher et al. (2001), where an empirically estimated constant is added to the standard deviation in the denominator.

4. The Bayes log posterior odds b_i introduced by Lönnstedt and Speed (2002).

The first two statistics (m and t) represent two common ways of dealing with the dispersion originating from replicate measurements. In microarray experiments with some thousand genes, we deal with a huge number of simultaneous measurements. Whereas the mean statistic m can be misleading for genes with high variances, the t–test statistic produces high outcomes for many genes with randomly low variances. It should be noted that in the case of small variances, a large test statistic is often caused by the denominator, whereas the amount of observed differential expression is small. Such, at best, small effects are usually of little practical relevance.

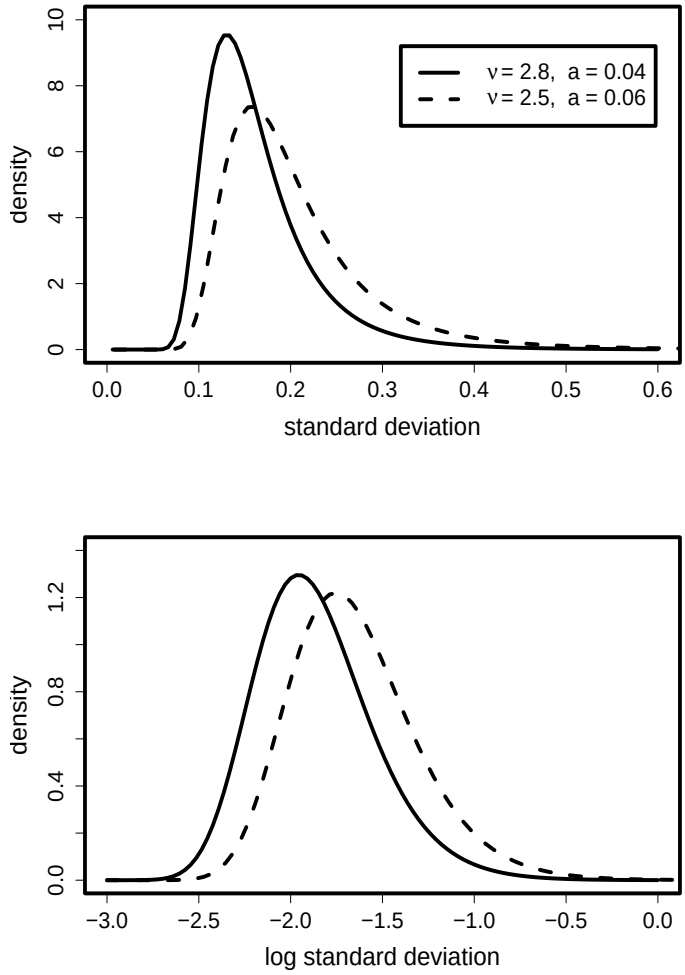

Figure 2: Densities of the standard deviations for a realistic parameter set (solid line) and for an unfavorable parameter set (dashed line).

The other two statistics are compromises between m and t, where the main principle is to add a constant to the standard deviation in the denominator. In the statistic s this constant is the 90%–percentile of the set of standard deviation estimates of all genes. In the statistic b this constant is derived from the parameters of the Empirical Bayes model, see Lönnstedt and Speed (2002), and then a monotone transformation is applied to this expression.

For the test statistics m, t and s we fitted a mixture of three normal components, $N(0, \sigma_0^2)$, $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, to the simulated data. The idea is that μ_1 and μ_2 specify typical levels of lower and higher gene expression, respectively. The maximum likelihood parameter estimates were obtained by using the EM-algorithm. Based on the estimated density we used the Bayes classifier specified in the previous section to assign every gene either to the group of differentially expressed genes or to the group of non-differentially expressed genes. For the test statistic b we used the theoretically optimal

log odds ratio of the Empirical Bayes model instead of the Bayes classifier from the fitted mixture model. For all statistics, we then compared these assignments with the predetermined group membership. Figures 3-7 illustrate the trade-off between false positives and false negatives. Figures 3-5 show results for data simulated from the normal mixture model, Figures 6-7 show results for data from the Bayes model of Lönnstedt and Speed. Figure 3 and Figure 4 provide a direct comparison for the two different parameter values for the inverse gamma distribution of the gene variances, with all other parameters held constant. Figure 4 is the only one with the more unfavorable values $\nu = 2.5$ and $a = 0.06$, in all other cases the true values are $\nu = 2.8$ and $a = 0.04$.

The ROC plots on the top of these figures show differences of the four test statistics in their ability to control the two error types. Values in the bottom left corner are preferred, since they represent low numbers of false decisions with respect to both differentially and non-differentially expressed genes. In the four bottom plots of Figure 3-7 white box plots represent false discoveries and black box plots false non-discoveries for a series of different costs. The cost values plotted on the abscissa are given by the series $\lambda_0 = 0.1, 0.2, \ldots, 0.8, 0.9, 0.95, 0.99, 0.999, 0.9999$. Without loss of generality, the costs λ_0 and $\lambda_1 = \lambda_2$ are normalized such that $\lambda_0 + 2\lambda_1 = 1$, hence it holds $\lambda_0 \in [0, 1]$. For each statistic, the relationship of costs (and importance) of the two types of false decisions determines the ratio of false discoveries and non-discoveries.

For a given data set, Figures 3-7 provide a guideline to choose cut-off points for the decision based on a given test statistic. It is well known in Bayesian statistics that the type one error decreases in λ_0, whereas the type two error is an increasing function in λ_0. This explains the trends in the box plot series of Figures 3-7. The point of intersection of the box plots for false discoveries and false non-discoveries depends on the actual number of differentially expressed genes in the microarray experiment. Depending on the goals of the experiment, a desirable value for λ may or may not be at the point of intersection.

If our chosen scenario provides a reasonably well representation of the actual data generating mechanism, then one might want to choose an acceptable point on the provided error graph. This point can be translated into costs that imply acceptance and rejection regions for the Bayes classifier and hence for the statistics. While the proposed models seem to provide realistic data sets, it should be checked of course whether the simulated data and the actual data look similar. A closer look at Figure 3 and Figure 4 shows that the recommendations for cut-off points are fortunately fairly robust with respect to gene variance distributions. The trends and cut-offs of the curves for false discoveries and false non-discoveries closely resemble each other, although the absolute numbers are significantly higher for the more unfavorable situation in Figure 4, which can also be seen from the ROC plots.

Based on the fitted mixture model and the Bayes classifier, estimates of the number of false discoveries are easily obtained. The accuracy of these estimates heavily depends on the true underlying data distribution. For the mixture model used in our simulations, where gene variances are drawn from an inverse gamma distribution, the estimates of the number of false discoveries were too optimistic, and the estimates of the number of false non-discoveries were too pessimistic. In a model with equal variances across genes these estimates are extremely accurate.

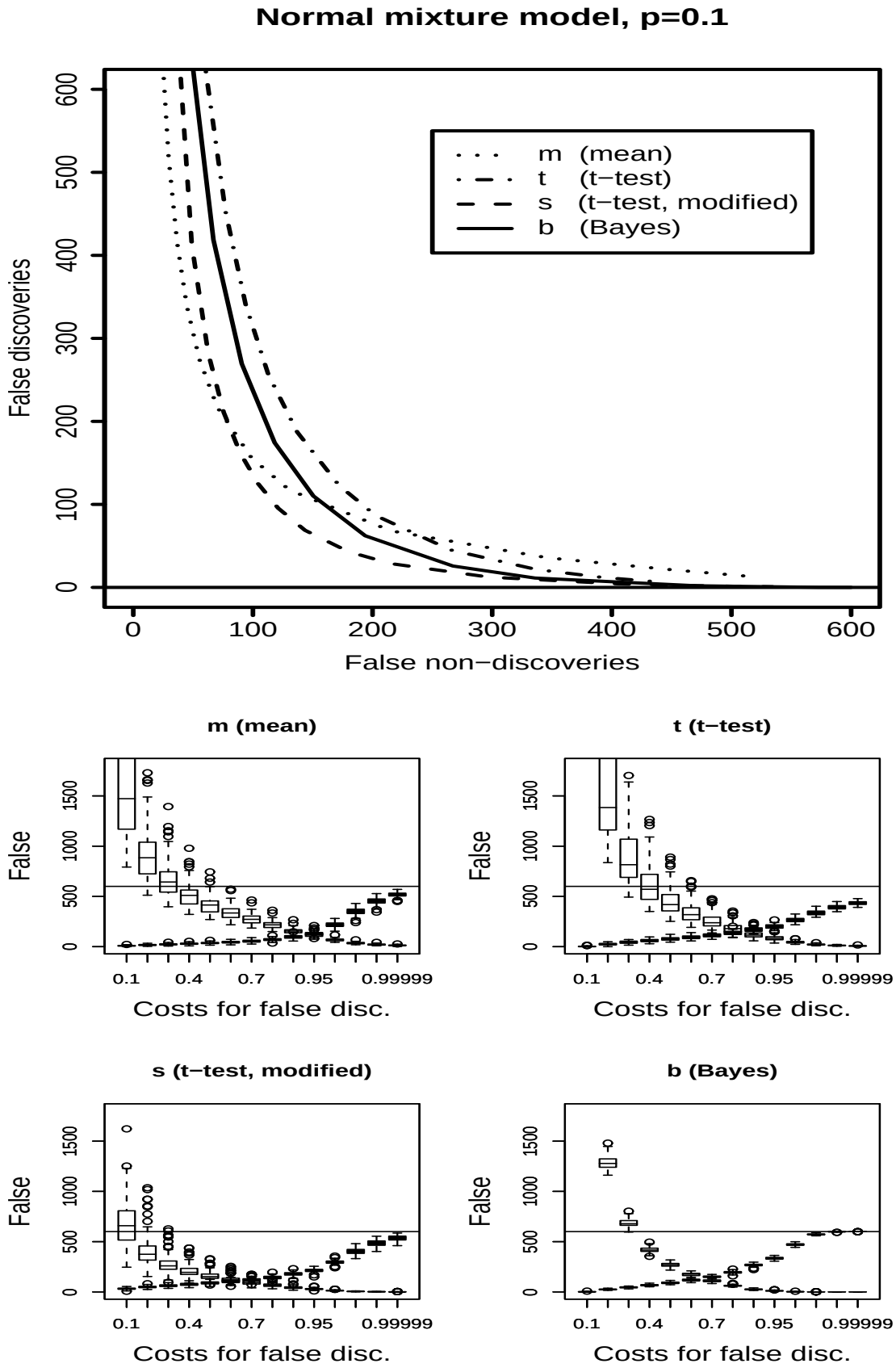

Figure 3: Mixture model, $p = 0.1$, false discoveries vs. false non-discoveries (top) and box plots for false decisions dependent on the cost λ_0 (bottom).

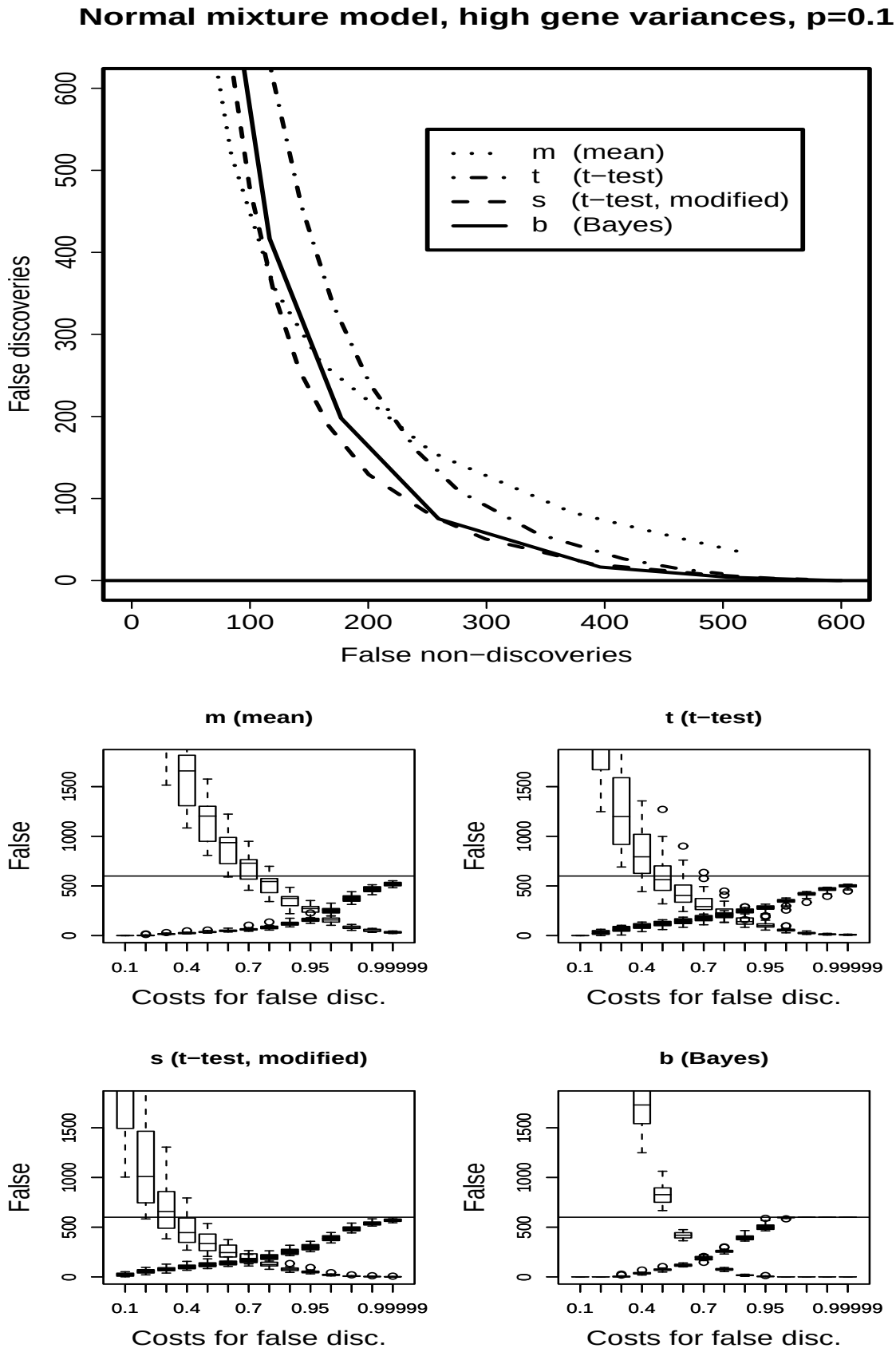

Figure 4: Mixture model, $p = 0.1$ as in Figure 3, with higher individual gene variances.

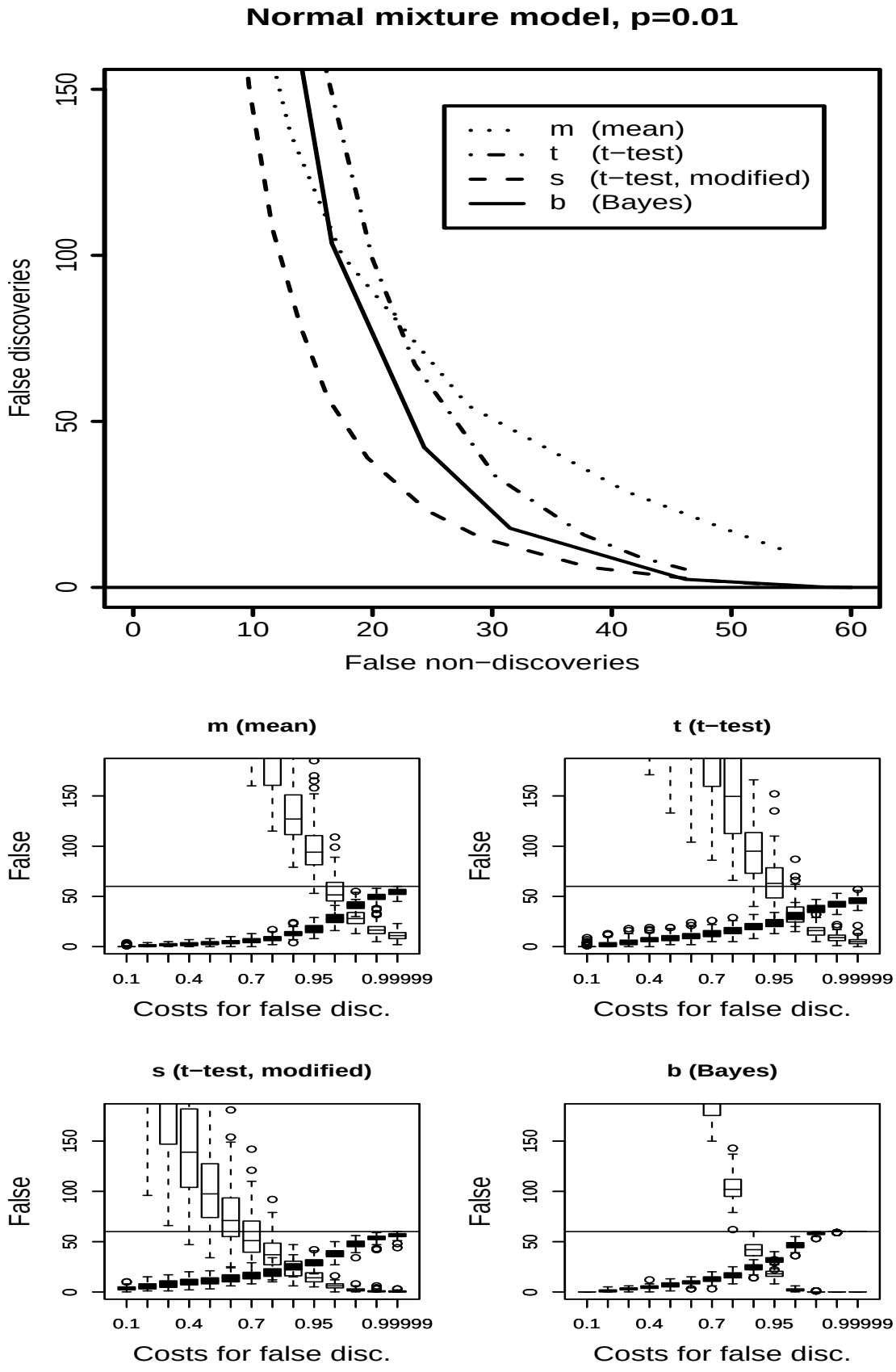

Figure 5: Mixture model, $p = 0.01$, false discoveries vs. false non-discoveries (top) and box plots for false decisions dependent on the cost λ_0 (bottom).

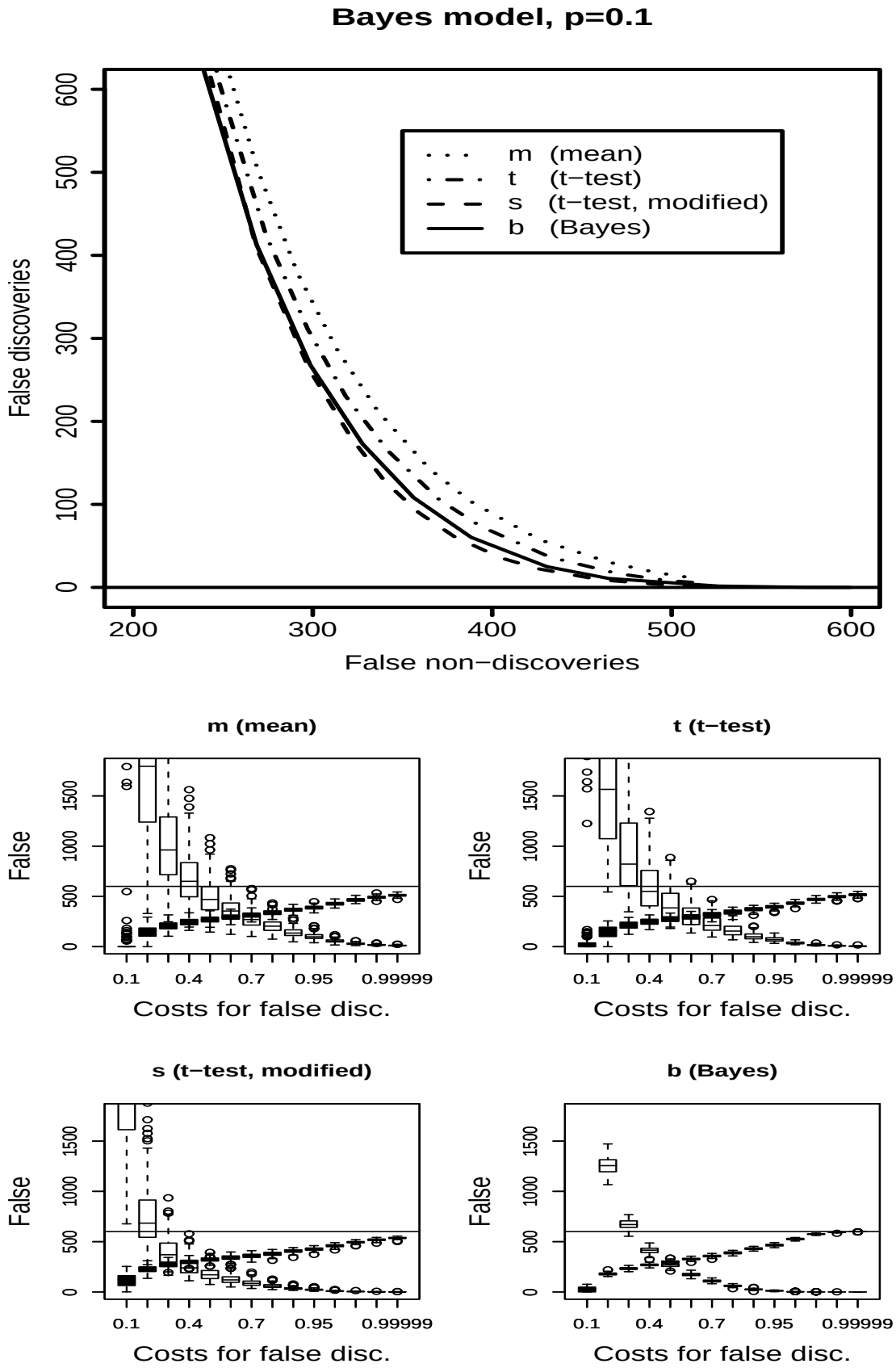

Figure 6: Bayes model, $p = 0.1$, false discoveries vs. false non-discoveries (top) and box plots for false decisions dependent on the cost λ_0 (bottom).

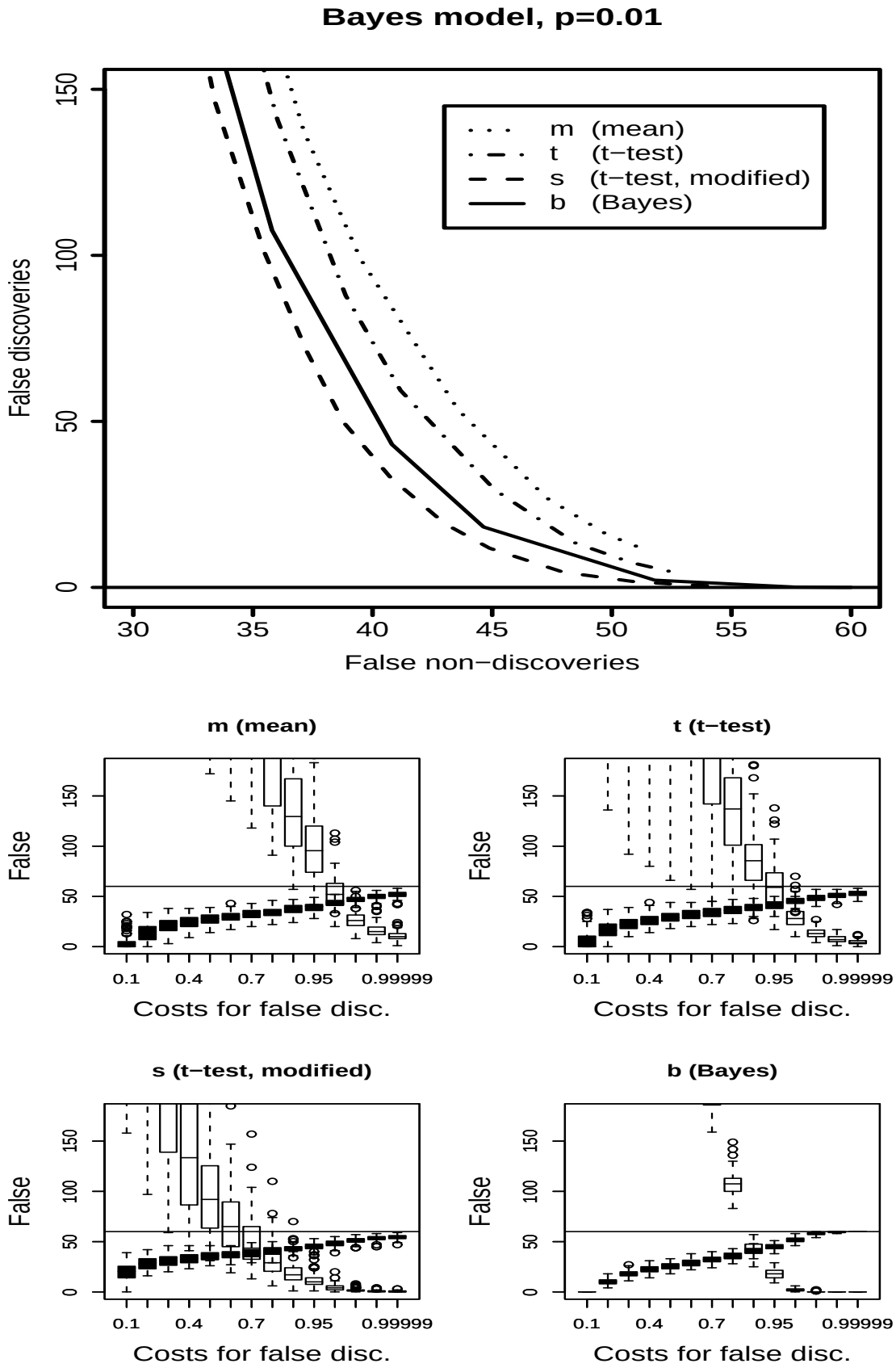

Figure 7: Bayes model, $p = 0.01$, false discoveries vs. false non-discoveries (top) and box plots for false decisions dependent on the cost λ_0 (bottom).

4 Discussion and Conclusions

We considered the problem of identifying differentially expressed genes for replicated microarray experiments. Our approach is to fit a normal mixture model to popular test statistics and then use a Bayes classifier with the estimated mixture components. This approach also allows us to estimate the number of false discoveries. Together with the analysis of realistic scenario models this provides guidelines for choosing appropriate decision boundaries for the involved test statistics.

We investigated our approach with simulated data sets. The simulation models have been chosen to provide data sets that are realistic according to previous experience with actual data. In these simulations it turned out that three mixture components often provide an already realistic representation of the data. Even though the implicit assumption of independence of the log ratios for different genes is unrealistic in most cases, our approach seems to provide some guidelines for a better informed use of test statistics available in the literature.

The ROC plots in Figures 3-7 provide a direct comparison of the four test statistics used in our simulations. The s statistic of Tusher et al. (2001) is the clear-cut winner for both models and for both a priori percentages of differentially expressed genes. This is even more striking, as the 90%-percentile added in the denominator of the statistic s was chosen arbitrarily and not further optimized by an additional adaptive step. The Berkeley statistic b generally produced the second-best results. Thus, as an overall conclusion, we can state that for microarray data with repeated measurements, test statistics that make some compromise between the mean m and the t–test statistic t are superior in minimizing false decisions regarding differential gene expression.

Our practical experience with microarray data tells us, that this is probably caused by the large number of low-intensity genes, which produce low means and low variances and thus too large t–statistics. In the models used for the simulations, the inverse gamma distribution for the standard deviation accounts for this fact. Apparently, there are significant differences with respect to optimal cut-offs and costs for 1% and for 10% differentially expressed genes. This indicates, that an a priori estimate for this percentage seems to be crucial, even when a suitable test statistic was chosen.

Acknowledgements

This work was supported by the “Deutsche Forschungsgemeinschafti” (JR, RA 870/2-2), by the “Fonds zur Förderung der wissenschaftlichen Forschung” (AF, J-1842 MAT) and through a consultancy on NIH grant R01HD037804-04 (JR, Claudia Kappen). We thank Andrea Krempler for carefully reading the manuscript and an anonymous referee for the valuable comments.

References

D. Bozinov and J. Rahnenführer. Unsupervised technique for robust target separation and analysis of dna microarray spots through adaptive pixel clustering. *Bioinformatics*, 18

(5):747–756, 2002.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer, Berlin, 1996.

S. Dudoit, Y.H. Yang, T.P. Speed, and M.J. Callow. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica*, 12(1):111–139, 2000.

B. Efron, R. Tibshirani, V. Goss, and G. Chu. Microarrays and their use in a comparative experiment. *J. Amer. Statist. Assoc.*, 96:1151–1160, 2001.

I. Lönnstedt and T. Speed. Replicated microarray data. *Statistica Sinica*, 12:31–46, 2002.

G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, and K.W. Tsui. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8:37–52, 2001.

W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 12:546–554, 2002.

V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98:5116–5124, 2001.

Y.H. Yang, M. Buckley, S. Dudoit, and T. Speed. Comparison of methods for image analysis on cdna microarray data. *Journal of Computational and Graphical Statistics*, 11:108–136, 2002.

Y.H. Yang, S. Dudoit, P. Luu, and T. Speed. Normalization for cdna microarray data. In *SPIE BiOS 2001*. San Jose, California, 2001.

Authors' addresses:

Dr. Jörg Rahnenführer
Max-Planck-Institut für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken
Germany

Tel. +49 681-9325-320
Fax +49 681-9325-399
Email: rahnenfj@mpi-sb.mpg.de
http://mpi-sb.mpg.de/˜rahnenfj

Dr. Andreas Futschik
University of Vienna,
Department of Statistics and Decision Support Systems
Universitätsstraße 5/9
A-1010 Vienna
Austria

Tel.: +43 1-4277-38634
Fax: +43 1-4277-38639
Email: andreas.futschik@univie.ac.at
http://mailbox.univie.ac.at/˜futscha3/