CHALMERS | GÖTEBORG UNIVERSITY

PREPRINT 2005:25

Weighted Analysis of Paired Microarray Experiments

ERIK KRISTIANSSON
ANDERS SJÖGREN
MATS RUDEMO
OLLE NERMAN

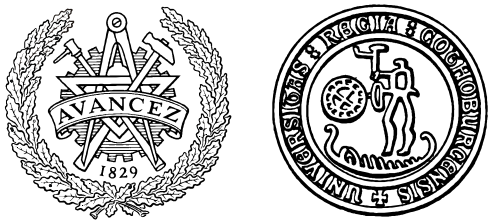
Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
GÖTEBORG UNIVERSITY
Göteborg Sweden 2005

Preprint 2005:25

Weighted Analysis of Paired Microarray Experiments

Erik Kristiansson
Anders Sjögren
Mats Rudemo
Olle Nerman

CHALMERS | GÖTEBORG UNIVERSITY



Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden
Göteborg, June 2005

Preprint 2005:25
ISSN 1652-9715

Matematiskt centrum
Göteborg 2005

Weighted Analysis of Paired Microarray Experiments

Erik Kristiansson[*], Anders Sjögren[*†],
Mats Rudemo and Olle Nerman

Abstract

In microarray experiments quality often varies, for example between samples and between arrays. The need for quality control is therefore strong. A statistical model and a corresponding analysis method is suggested for experiments with pairing, including designs with individuals observed before and after treatment and many experiments with two-colour spotted arrays. The model is of mixed type with some parameters estimated by an empirical Bayes method. Differences in quality are modelled by individual variances and correlations between repetitions. The method is applied to three real and several simulated datasets. Two of the real datasets are of Affymetrix type with patients profiled before and after treatment, and the third dataset is of two-colour spotted cDNA type. In all cases, the patients or arrays had different estimated variances, leading to distinctly unequal weights in the analysis. We suggest also plots which illustrate the variances and correlations that affect the weights computed by our analysis method. For simulated data the improvement relative to previously published methods without weighting is shown to be substantial.

Keywords

Quality control, QC, Quality Assurance, QA, Quality Assessment, Empirical Bayes, DNA Microarray

[*]both authors contributed equally, order was randomised
[†]corresponding author; e-mail: anders.sjogren@math.chalmers.se

1 Introduction

DNA microarrays are strikingly efficient tools for analysing gene expression for large sets of genes simultaneously. They are often used to identify genes that are differentially expressed between two conditions, e.g. before and after some treatment. A drawback is that the technology involves several consecutive steps, each exhibiting large quality variation. Thus there is a strong need for quality assessment and quality control to handle occurrences of poor quality, as is clearly pointed out in Johnson and Lin (2003) and Shi et al. (2004).

Despite the observed need for effective quality control, standard operating procedures for quality assurance of the entire chain of processing steps have only recently been proposed (Ryan et al., 2004, for one-channel experiments). However, even utilising an optimal quality control procedure aiming at removing low quality arrays and/or individual gene measurements (e.g. spots), there will always be a marginal region with some measurements being of decreased quality without being worthless, as noted in Ryan et al. (2004). Consequently, it should be possible to make progress by integrating quality control quantitatively into the analysis following the lab steps and low-level analysis, taking quality variations into account.

When integrating the quality concept into the analysis, the quality of different parts of the dataset should ideally be estimated from data and used in the subsequent selection of differentially expressed genes. Here we introduce a method, called *Weighted Analysis of paired Microarray Experiments* (referred to as WAME), for the analysis of paired microarray experiments, e.g. comparison of pairs of treatment conditions and most two-colour experiments. WAME aims at estimating array-wide quality deviations and integrates the quality estimates in the statistical analysis. Only the observed gene expression ratios are used in the quality assessment, making the method applicable to most paired microarray experiments, independent of which DNA microarray technology is used.

In short WAME identifies and downweights repetitions (biological or technical) of pairs (corresponding to individuals or to arrays) with decreased quality for many genes. Repetitions with positively correlated variations, e.g. caused by shared sources of variation, are similarly down-weighted. Thus, estimates of differential expression with improved precision and tests with increased power are provided.

As a useful complement to the WAME analyses we suggest pair-wise plots of log-ratios of gene expression measurements. Such plots are supplied for all three real datasets analysed, and they are particularly useful for under-

standing which patients or arrays that are up- or downweighted.

In the adopted model, log ratios of measured RNA-levels are assumed normally distributed. The covariance structure is specified by parameters of two types: (i) a global covariance matrix signifying different quality for different repetitions and (ii) gene specific multiplicative factors. The latter have inverse gamma prior distribution with one gene-specific parameter, which is estimated by an empirical Bayes method.

The paper is organised as follows. In the next section, a background and a selection of previous work in the field are presented. This is followed by a detailed description of our model. Methods for estimating the parameters and a likelihood ratio test for identifying differentially expressed genes are derived. In the following section simulations are used to compare WAME to four currently used methods: (i) average fold change, (ii) ordinary t-test, (iii) the penalized t-statistic of Efron et al. (2001), and (iv) the moderated t-statistic of Smyth (2004). Next, WAME is applied to three real datasets, the *Cardiac* dataset of Hall et al. (2004), the *Polyp* dataset of Benson et al. (2004) and the *Swirl* dataset (Dudoit and Yang, 2003). The results obtained are discussed in a subsequent section and some derivations and mathematical details are given in an appendix.

2 Background

To put the quality control aspect of our model into context, the different steps and sources of variation in typical paired microarray experiments are outlined below. In addition, a selection of publications dealing with quality control for microarray experiments are briefly reviewed.

2.1 Sources of variation in typical microarray experiments

The first step, after decision on experimental design, of a microarray experiment aiming at identifying differentially expressed genes would typically be to determine how biological samples should be acquired. In experiments dealing with homogeneous groups of single cell organisms, such as yeast, in highly controlled environments, this task is typically less complex than when dealing with heterogeneous groups of multicellular organisms, such as humans. Here selection of subjects and cells from the relevant organ, e.g. by biopsy or laser dissection, are complicated tasks.

From the biological sample the following lab-steps are performed: RNA extraction, reverse transcription (and *in vitro* transcription), labelling, hybridisation to arrays and scanning. The parts of the scanned images corresponding to the different genes (i.e. spots or probe-pairs) are identified and quantified. In addition, background correction may be performed. Subsequently, normalisation of the quantified measurements is performed to handle global differences. In the case of Affymetrix type arrays, 11-20 pairs of quantitative measurements are combined into one expression level estimate for each gene. For an experiment of paired type, one $\log_2$-ratio of the expression level estimates is calculated for each pair and gene. These $\log_2$-ratios are then used to examine which genes are differentially expressed.

In several of the steps mentioned above there are substantial quality variations. For example, the quantity and quality of the RNA in biopsies may vary considerably. There are sometimes evidence of poor quality making it possible to remove obviously worthless samples. Nevertheless, there will always be a marginal region with measurements of reduced quality without being worthless. In addition, some variations are hard to detect before the actual normalised $\log_2$-ratios are computed, e.g. the representativeness in tissue distribution of human biopsies. An additional aspect of quality control is systematic errors, where the variations of different repetitions are correlated. This could be due to shared sources of variation, such as simultaneous processing in lab steps or non-representative tissue composition in the biopsies.

Another potentially important factor is the quality of the arrays used for the measurements. Flaws in the manufacturing process might make measurements for individual genes inferior. This is more of a problem in the case of spotted arrays, for which there are only one or a few spots per gene. However, such bad spots can often be detected. The quality control in the actual manufacturing of microarrays is certainly very important but will not be further discussed here.

2.2 A brief review of some relevant publications

In Johnson and Lin (2003) and Shi et al. (2004) the general need for improved quality assurance in the context of DNA microarray analysis is emphasised. Tong et al. (2004) implement a public microarray data and analysis software and note that "Although the importance of quality control (QC) is generally understood, there is little QC practise in the existing microarray databases". They include some available measures of quality for different steps in the analysis in their database.

Dumur et al. (2004) survey quality control criteria for the wet lab steps of Affymetrix arrays, going from RNA to cDNA. Additionally, three sources of technical variation (hybridisation day, fluidic scan station, fresh or frozen cDNA) are evaluated using an ANOVA model.

Ryan et al. (2004) present guidelines for quality assurance of Affymetrix based microarray studies, utilising a variety of techniques for the different steps, some of which are shown to agree. A sample quality control flow diagram is suggested, including steps from extracted RNA to the quantified arrays.

Chen (2004) aims at screening out ineligible arrays (Affymetrix type), using a graphical approach, so called *2D image plots*, to display grouped data. Park et al. (2005) similarly aim at identifying outlying slides in two-channel experiments by using scatterplots of transformed versions of the signals from the two channels.

Tomita et al. (2004) use correlation between arrays (Affymetrix type) to evaluate the RNA integrity of the individual arrays, by forming an average correlation index (ACI). The ACI is shown to correlate with several existing quality factors, such as the 3'-5' ratio of GAPDH.

Several papers have been written on the quality control of individual measurements of genes (spots or probes). Wang et al. (2001, 2003) define a spot-wise composite score from various quantitative measures of quality of individual spots in spotted microarrays. They further perform evaluations on several in-house datasets, showing that when bad spots are removed, the variance of all gene-specific ratios in one chip is decreased. In Hautaniemi et al. (2003) Bayesian networks are used to discriminate between good and bad spots with training data provided by letting experienced microarray users examine the arrays by hand.

In the papers discussed above the countermeasure against low-quality spots or arrays is to treat them as outliers and to remove them. Again, there will always be a marginal region with some measurements being of decreased quality without being worthless. An interesting approach using weighted analysis of the microarray gene expression data is due to Bakewell and Wit (2005). The starting point is a variance component model for the log expression mean for a spot i with variance $\sigma_b^2 + \sigma_i^2/m_i$, where σ_b^2 is the variance between spots while σ_i^2 is the variance between pixels in spot i with the effective number m_i of pixels. For each gene the spots are weighted inversely proportional to estimated variances, and different genes are essentially treated independent of each other. Only quality deviations of the actual hybridised spots are included in the model.

In Yang et al. (2002) the variance of different print tip groups or arrays

in cDNA experiments are estimated by a robust method. The need for scale normalisation between slides is determined empirically, e.g. by displaying box plots for the log ratios of the slides.

The model we propose (WAME) assesses the quality of different arrays quantitatively by examining the computed $\log_2$-ratios. Thus, quality deviations in all steps leading to the gene expression estimates are included, as long as the quality deviations occur for a wide variety of measured genes. Furthermore, shared systematic errors are taken care of via estimated covariances between repetitions. The assessed qualities are incorporated into the analysis based on the statistical model presented in the next sections.

In microarray experiments there are often relatively few replicates, resulting in highly variable gene-specific variance estimates. To use the information in the large number of measured genes to handle this problem, an empirical Bayes approach (Robbins, 1956; Maritz, 1970) can be taken, determining a prior distribution from the data, thus moderating extreme estimates. This approach has been used in Baldi and Long (2001), Lönnstedt and Speed (2002) and Smyth (2004).

3 The model

The experimental layouts studied in the present paper are restricted to comparisons of paired observations from two conditions . For each gene $g = 1, \ldots, N_G$ and each pair of measurements $i = 1, \ldots, N_I$, let X_{gi} with expected value μ_g be the normalised $\log_2$-ratio of the observed gene expressions from the two conditions. Thus, μ_g measures the expected $\log_2$ ratio of the RNA concentrations of the two conditions.

In Section 2.1 it was noted that there may exist dependencies between repetitions, e.g. due to systematic errors. Furthermore, different arrays may have different precision in their measurements of the gene expressions. To describe this, we use a covariance structure matrix Σ which models precision as individual variances for the different repetitions and dependencies between repetitions as covariances.

Due to both technical and biological reasons the observations for the different genes have different variability, and a gene-specific multiplicative factor c_g for the covariance matrix is introduced. The c_g-variables for different genes are assumed to be independent. Given c_g the vector $\mathbf{X_g}$ consisting of all repetitions for gene g is assumed to have a N_I-dimensional normal distribution with mean vector $\mu_g\mathbf{1}$ and covariance matrix $c_g\Sigma$. The vectors $\mathbf{X_g}$ for different genes are also assumed independent. This independence

assumption is optimistic but we believe that it is not critical in the sigma-estimation step owing to the large number of genes.

In microarray experiments, the number of experimental units is typically fairly small and estimates of c_g utilising only information from the measurements with gene g may be highly variable. Therefore prior information is introduced as a prior distribution for c_g, which serves to moderate the estimates of c_g. The prior for c_g is assumed to be an inverse gamma distribution with a parameter α determining the spread of the distribution, in effect determining the information content in the prior. The inverse gamma distribution is a conjugate prior distribution for the variance of a normal distribution and has as such been used in Bayesian and empirical Bayesian analysis of microarray data before (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Smyth, 2004).

The model can be summarised as follows: We observe $\mathbf{X}_g = (X_{g1}, \ldots, X_{gN_I})$ where $g = 1, \ldots, N_G$. Let Σ be a covariance matrix with N_I rows and columns, c_g a set of gene-specific variance scaling factors and α a hyperparameter determining the spread of the prior distribution for c_g. Then for fixed μ_g, Σ and α,

$$\begin{aligned} c_g &\sim \Gamma^{-1}(\alpha, 1) \quad \text{and} \\ \mathbf{X}_g \mid c_g &\sim \mathbf{N}_{N_I}\left(\mu_g \mathbf{1}, c_g \Sigma\right), \end{aligned} \tag{1}$$

and all variables corresponding to different genes are assumed independent.

4 Inference

4.1 Estimation of a scaled version of the matrix Σ

Estimating Σ may appear easy but it turns out to be rather intricate and there are several issues involved.

Firstly, there are trivial solutions that give infinite likelihood of the model. Just put one gene-specific mean value μ_g equal to the observation of one of the repetitions and the corresponding variance equal to zero. To avoid this complication the assumption that the differential expression of most genes is approximately zero is introduced temporarily. This assumption is not as consequential as it might sound, since it is made by most of the procedures that have become *de facto* standard in the (preceding) normalisation step, one example being the loess normalisation method (Yang et al., 2002). Nevertheless, it does limit the set of experimental setups that can be treated and the proportion of genes that are regulated must not be too large. The

impact of this assumption is further investigated by the simulation study in Section 5.2. For the rest of Section 4.1, μ_g is thus set equal to zero for all $g = 1, \ldots, N_G$.

Another issue is the scaling of Σ. For each gene, the covariance matrix is scaled with the random variable c_g which has an inverse gamma distribution with a parameter which is unknown in a first stage. To address this issue, the estimation of Σ is performed in two steps. In the first step, a transformation is applied to $\mathbf{X}_g$ such that the transformed vector has a distribution that is independent of c_g. To simplify notation the index g will be dropped from $\mathbf{X}_g$ and c_g in the rest of this section. Let $\mathbf{U} = (U_1, \ldots, U_{N_I})$ where

$$U_i = \begin{cases} X_1 & \text{if } i = 1 \\ X_i/X_1 & \text{if } 2 \leq i \leq N_I. \end{cases}$$

The distribution of the vector $\mathbf{U}$ has the density

$$f_{\mathbf{U} \mid c,\Sigma}(\mathbf{u}) = f_{\mathbf{X} \mid c,\Sigma}(\mathbf{x}(\mathbf{u}))|J(\mathbf{u})|$$

where J is the corresponding Jacobian. Some algebra shows that the scaling factor c cancels for $U_2, \ldots, U_{N_I}$ and by integrating over U_1, we get the density

$$\begin{aligned} f_{U_2,\ldots,U_{N_I} \mid \Sigma}(u_2, \ldots, u_{N_I}) &= \int_{-\infty}^{\infty} f_{\mathbf{U} \mid c,\Sigma}(\mathbf{u})\, du_1 \\ &= C\, |\Sigma|^{-1/2} \left[v^{\mathrm{T}} \Sigma^{-1} v\right]^{-N_I/2}, \end{aligned} \tag{2}$$

where C is a normalisation constant and $v = (1, u_2, \ldots, u_{N_I})$. The distribution (2) is independent of c and the marginal distribution of u_i is a Cauchy distribution translated with $\rho_{1,i}\sigma_{i,i}/\sigma_{1,1}$ and scaled with $\sqrt{1-\rho_{1,i}^2}\sigma_{i,i}/\sigma_{1,1}$, where $\rho_{1,i}$ is the correlation between X_1 and X_i and $\sigma_{i,i}$ is the variance of X_i. This shows that $\rho_{1,i}$ and $\sigma_{i,i}/\sigma_{1,1}$ are identifiable. Analogously, from the one dimensional Cauchy distributions of $U_j/U_k = X_j/X_k$, $j = 2, \ldots, N_I$ and $k = 2, \ldots, N_I$ it follows that all other correlations and variance ratios are identifiable as well.

From (2) we see that the distribution of $(U_2, \ldots, U_{N_I})$ is unchanged if we multiply Σ with a constant. Let us therefore fix one element of Σ, e.g. we set the first element in the first row equal to one. Let Σ^* denote the matrix thus obtained. Then

$$\Sigma^* = \lambda\Sigma\,, \tag{3}$$

and the constant λ will be estimated together with the hyperparameter α as described below in Section 4.2. Thus estimation of the covariance matrix Σ

will be carried out in two steps: first estimate Σ^* with one element fixed and then estimate λ.

Numerical maximum likelihood based on the distribution (2) is used to produce a point estimate of Σ^*. The computational complexity grows as N_I^2 since the number of unknown parameters $N_I(N_I+1)/2$. To get an efficient implementation C/C++ is combined with R (R Development Core Team, 2004). The resulting computational time for three arrays is less than a second and for 30 arrays it takes a few hours.

4.2 Estimation of the hyperparameter α and the scale λ

In this section, we develop methods for estimation of the hyperparameter α as well as the scale parameter λ in (3). From the model assumptions in Section 3 we recall that c_g has an inverse gamma distribution with hyperparameter α, e.g.

$$c_g \mid \alpha \sim \Gamma^{-1}(\alpha, 1).$$

The inference of α will be based on the statistic

$$S_g = (A\mathbf{X}_g)^{\mathrm{T}}(A\Sigma A^{\mathrm{T}})^{-1}A\mathbf{X}_g,$$

where A is an arbitrary $N_I - 1 \times N_I$ matrix with full rank and each row sum equal to 0. It follows that the distribution of S_g conditioned on c_g is a scaled chi-square distribution with $N_I - 1$ degrees of freedom,

$$S_g \mid c_g \sim c_g \cdot \chi^2_{N_I-1}.$$

The unconditional distribution of S_g can be calculated by use of the fact that a gamma distribution divided by another gamma distribution has an analytically known distribution, a beta prime distribution (Johnson et al., 1995, page 248). Thus,

$$S_g \mid \alpha \sim 2 \times \beta'\left((N_I-1)/2, \alpha\right),$$

which has the density function

$$f_{S_g \mid \alpha}(s_g) = \frac{1}{2}\frac{\Gamma(\alpha + (N_I-1)/2)}{\Gamma(\alpha)\Gamma((N_I-1)/2)}\frac{(s_g/2)^{(N_I-1)/2-1}}{[1+s_g/2]^{\alpha+(N_I-1)/2}}.$$

In the same fashion, denote the variance estimator based on Σ^* in (3) by S_g^*, that is,

$$S_g^* = (AX_g)^{\mathrm{T}}(A\Sigma^* A^{\mathrm{T}})^{-1}AX_g\ .$$

It follows that, $S_g^* = S_g/\lambda$ so

$$S_g^* \mid \alpha, \lambda \sim 2/\lambda \times \beta' \left((N_I - 1)/2, \alpha\right) \ .$$

Assuming independence between the genes, α and λ can now be estimated by numerical maximum likelihood. The estimated value of the (unscaled) covariance matrix Σ can then be calculated from Equation (3). Results from simulations show that the estimation of α and λ is accurate enough for realistic values (results not shown). In the following sections, these parameters are therefore assumed to be known.

4.3 The posterior distribution of c_g

The posterior distribution of c_g is not explicitly used in the calculations above, but still of general interest. As previously mentioned, the distribution of S_g conditioned on c_g is a scaled chi-square distribution with $N_I - 1$ degrees of freedom. Since chi-square distributions and inverse gamma distributions are conjugates, the posterior distribution of c_g given S_g is an inverse gamma distribution as well. We find

$$c_g \sim \Gamma^{-1}\left(\alpha, 1\right)$$

$$c_g \mid S_g \sim \Gamma^{-1}\left(\alpha + (N_I - 1)/2, 1 + \frac{S_g}{2}\right) \ ,$$

and the prior can be interpreted as representing 2α pseudo observations, which add a common variance estimate to all genes. A discussion regarding the use of this model in microarray analysis can be found in (Lönnstedt and Speed, 2002) and (Smyth, 2004) and a general discussion in (Robert, 2003, Section 4.4).

4.4 Inference about μ_g

In this section we derive a statistical test for differential expression based on the WAME model. The hypotheses for gene g can be formulated as

$$\begin{aligned} H_0 &: \text{ gene } g \text{ is not regulated } (\mu_g = 0) \\ H_A &: \text{ gene } g \text{ is regulated } (\mu_g \neq 0). \end{aligned}$$

A test suitable for the hypothesis H_0 is the likelihood ratio test (LRT) based on the ratio of the maximum values of the likelihood function under the

different hypotheses. With our notation we reject H if

$$\frac{\sup\limits_{H_A} L\left(\mu_g|\mathbf{x}_g\right)}{\sup\limits_{H_0} L\left(\mu_g|\mathbf{x}_g\right)} = \frac{\sup\limits_{\mu_g \neq 0} L\left(\mu_g|\mathbf{x}_g\right)}{L\left(0|\mathbf{x}_g\right)} \geq k, \tag{4}$$

where k, $1 \leq k < \infty$, sets the level of the test. To calculate the likelihood function, we need to integrate over c_g, e.g.,

$$\begin{aligned} L\left(\mu_g|\mathbf{x}\right) &= \int f_{\mathbf{x} \mid \mu_g, c_g, \Sigma}(\mathbf{x}) f_{c_g \mid \alpha}(c_g)\, dc_g \\ &= (2\pi)^{-N_I/2} \left|\Sigma\right|^{-1/2} \frac{\Gamma(N_I/2+\alpha)}{\Gamma(\alpha)} \left[\frac{\left(\mathbf{x}_g - \mu_g \mathbf{1}\right)^{\mathrm{T}} \Sigma^{-1} \left(\mathbf{x}_g - \mu_g \mathbf{1}\right)}{2} + 1\right]^{-(\alpha+N_I/2)}. \end{aligned}$$

It is now straight forward to calculate the denominator $L(0|\mathbf{x}_g)$ in (4) and some algebra shows that the numerator is maximised by $\hat{\mu}_g = \bar{x}_g^w$, where

$$\bar{x}_g^w = \frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\mathbf{x}_g \ ,$$

is a *weighted mean value* of the observations. Analogously, we define the random variable $\bar{X}_g^w$ by replacing $\mathbf{x}_g$ with $\mathbf{X}_g$. Then,

$$\bar{X}_g^w | c_g \sim \mathbf{N}\left(\mu_g, \frac{c_g}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right)$$

and it can be shown that

$$\mathbf{w}^{\mathrm{T}} = \frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}} \tag{5}$$

is the weight vector that minimises the variance of $\mathbf{w}^{\mathrm{T}}\mathbf{X}_g$. The weights in equation (5) will depend on the covariance matrix as follows. A repetition with high variance will have a low weight while a repetition with low variance will have a high weight. Moreover, a positive high correlation between repetitions will cause decreased weights. Note that if a repetition is highly correlated with a repetition with lower variance, its weight can actually become negative. According to the theory, this is nothing strange but practically this is of course not satisfying. Fortunately, such extreme cases seem to be rare in the microarray context and if they appear, the source of the correlation should be investigated and one could consider removing the negatively weighted repetition.

Evaluation of the likelihood function at 0 and $\bar{x}_g^w$ and a few calculations show that the inequality (4) is equivalent to

$$\frac{|\bar{x}_g^w|}{\sqrt{s_g + 2}} \geq k'$$

where s_g is the observed value of S_g defined in Section 4.2 and k' is some non-negative constant. Thus if we define the statistic T_g as

$$T_g = \sqrt{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}\,(N_I - 1 + 2\alpha)}\;\frac{\bar{X}_g^w}{\sqrt{S_g + 2}}$$

and the null hypothesis is rejected if

$$|T_g| \geq k'',$$

where k'' is another non-negative constant. The statistic T_g will be referred to as the *weighted moderated t-statistic* since it is a weighted generalisation of the moderated t-statistic derived by Lönnstedt and Speed (2002) and refined by Smyth (2004). Indeed, if all repetitions have equal estimated variances and no estimated correlations, T_g becomes equivalent to the result in Section 3 in Smyth (2004). To calculate the value of k'' that corresponds to a given level of the test, the distribution of T_g needs to be derived. Under the null hypothesis, it turns out to be a t-distribution with $2\alpha + N_I - 1$ degrees of freedom,

$$T_g \;\sim\; t_{2\alpha+N_I-1}.$$

5 Results from simulations

5.1 Comparison to similar gene ranking methods

A simulation study was done to compare the performance of WAME to four published methods. These methods were

- Average fold-change
- Ordinary t-statistic
- Efron's penalized t-statistic
- Smyth's moderated t-statistic

The average fold-change for a gene is simply the mean value over all the observed $\log_2$-ratios and the ordinary t-statistic is the average fold-change divided by the corresponding sample standard deviation. These two methods have traditionally been popular gene ranking methods and it is therefore interesting to see how they perform. Another method introduced in (Efron et al., 2001) is the penalized t-statistic which is a modified version of the

ordinary t-statistic where a constant has been added to the sample standard deviation. The motivation for this adjustment is the unreliability of the t-statistic in situations when only a few repetitions are used. The constant used here was chosen as the 90th percentile of the empirical distribution of the sample standard deviations, according to Efron et al. (2001). Finally, the moderated t-statistic is included. It was developed and implemented by Smyth (2004) and it is available in the R package LIMMA (Smyth et al., 2003). The moderated t-statistic can be seen as a refined version of the B-statistic which was first presented in Lönnstedt and Speed (2002). In the paired microarray context, WAME is a generalisation of LIMMA in the sense that the two models are identical when all repetitions have the same variance and no correlations exist.

All methods were applied to a series of simulated datasets with different settings and the number of true positives as a function of false positives was plotted, generating several so called receiver operating characteristic (ROC) curves. The average over 100 datasets was used to produce a single curve where each dataset was created as follows. The number of genes (N_G) was fixed to 10000, the number of repetitions (N_I) to 4 and the hyperparameter α to 2. These values were chosen since they are typical for real datasets. The covariance matrix Σ is fixed and for each gene g the following steps were done.

1. c_g was sampled from an inverse gamma distribution according to the model specification.

2. A vector of $N_I = 4$ independent observations was drawn from a normal distribution with man value zero and variance one. This vector was then multiplied by the square-root matrix of Σ.

3. If this particular gene was selected to be regulated, then the absolute mean value for each of the N_I elements was drawn from a uniform distribution between 0 and 2.

5% of the genes were randomly selected and set to be upregulated. Analogously, 5% were downregulated resulting in totally 10% regulated genes. It should be noted that it is only the total number of regulated genes that had an impact on the performance for the different methods, not the number of upregulated genes compared to the number of downregulated genes.

Four cases, all with different covariance matrices, were studied. In the first case, all of the repetitions had variances equal to 1 and there were no correlations, thus Σ was an identity matrix. The ROC curves produced by

the simulated data can be seen in the upper part of Figure 1. WAME and LIMMA performs best, closely followed by the penalized t-statistic. Note that WAME and LIMMA have almost identical performance in this case and, as mention above, this was expected since the weighted moderated t-statistic and the moderated t-statistic are almost equivalent for this setting. Another interesting detail is the weak performance of the t-statistic due to its instability issues when only few repetitions are used.

In the second case, different variances were introduced. Σ was again a diagonal matrix but with the values 0.5, 1, 1.5 and 2 on the diagonal, thus all correlations were again zero. The ROC curves can be seen in the lower part of Figure 1. As before, WAME and LIMMA are the methods that performs best, but in this case, WAME performs better since it put less weight on the repetitions with high variance.

To investigate the impact of correlations, the third case used

$$\Sigma = \begin{pmatrix} 1.0 & 0.4 & 0.2 & 0.0 \\ 0.4 & 1.0 & 0.4 & 0.2 \\ 0.2 & 0.4 & 1.0 & 0.4 \\ 0.0 & 0.2 & 0.4 & 1.0 \end{pmatrix}. \tag{6}$$

This corresponds to a case when there are both high and low correlations between the repetitions. The upper part of Figure 2 shows that WAME performs slightly better than both LIMMA and the penalized t-statistic since it estimates the correlations and takes them into account.

Finally, in the fourth case both different variances and correlations were included. The variances and correlations were identical to the ones in the second and third cases respectively, i.e. variances of 0.5, 1.0, 1.5, 2.0 and correlations of 0, 0.2 and 0.4, the latter placed according to (6). The result can be seen in the lower part of Figure 2. Here, the largest advantage of using WAME can be seen. For a rejection threshold such that half of the selected genes are true positives, using WAME results in almost a third less false positives which can correspond to hundreds of genes.

All four simulations show that WAME and its weighted moderated t-statistic perform at least as good as the moderated and penalized t-statistics. In the case of both different variances and correlations between the repetitions, WAME performs clearly better than all of the included methods. Both the average fold-change and the ordinary t-statistic have poor performance in the current setting with only four repetitions.

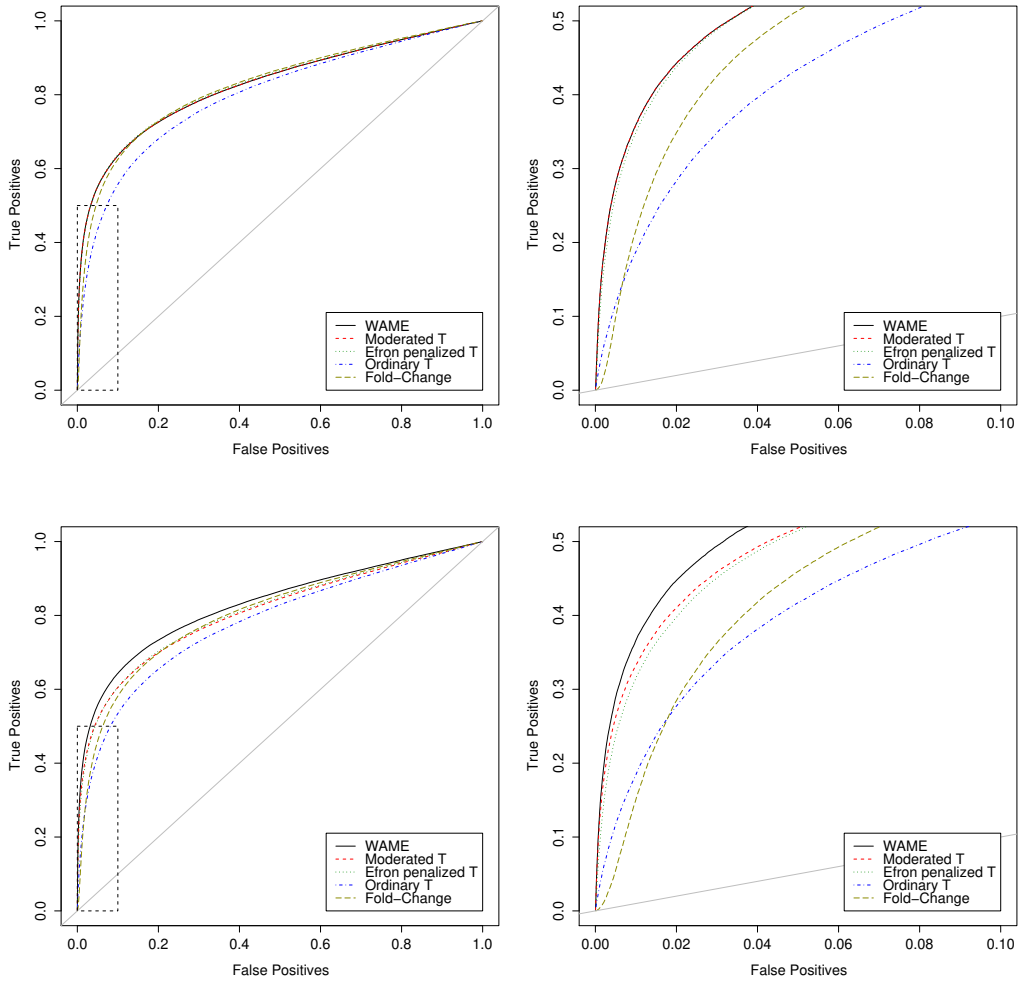


Figure 1: ROC curves from simulated data. The pair at the top, from the first case, show the performance of the evaluated methods on data with equal variances of 1 for all replicates and no correlations. The pair at the bottom, from the second case, analogously show the performance on data with different variances of 0.5, 1, 1.5, 2 and no correlations. The parameters used for these two simulations were as follows. $N_G = 10000$, $N_I = 4$, $\alpha = 2$ and 10% of the genes were regulated. The figures to the right are magnifications of the dashed boxes to the left.

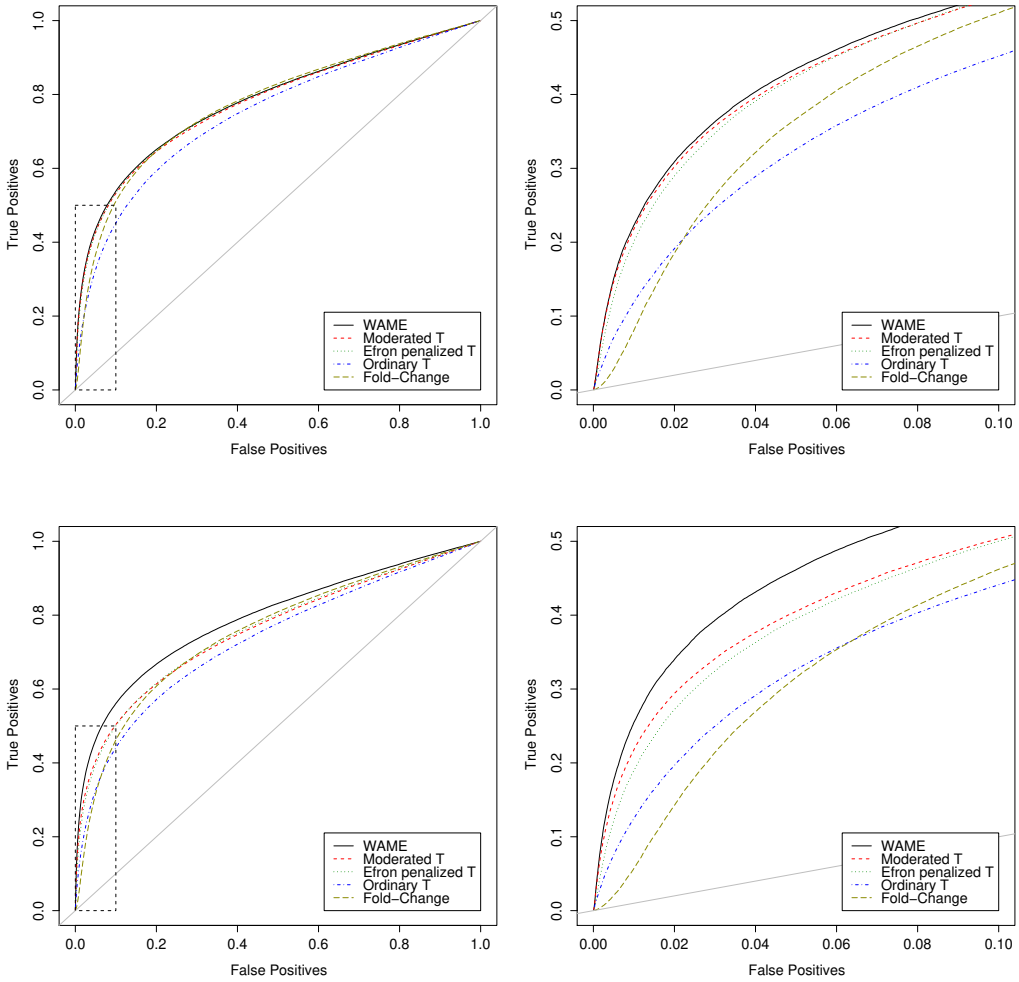


Figure 2: ROC curves from simulated data. The pair at the top, from the third case, show the performance of the evaluated methods on data with equal variances of 1 for all replicates and correlations of 0, 0.2 and 0.4, placed according to (6). The pair at the bottom, from the fourth case, analogously show the performance on data with different variances of 0.5, 1, 1.5, 2 and correlations of 0, 0.2 and 0.4, placed according to (6). The parameters used for these two simulations were as follows. $N_G = 10000$, $N_I = 4$, $\alpha = 2$ and 10% of the genes were regulated. The figures to the right are magnifications of the dashed boxes to the left.

5.2 Evaluation of the point estimation of Σ

The estimation of Σ is one of the crucial steps when applying WAME since errors made will affect estimates of other entities such as α and the weighted mean value $\bar{x}_g^w$. The resulting precision and accuracy when numerical maximum likelihood is applied to the distribution in equation (2) are therefore interesting questions, both when the model assumptions hold and when they are violated. In an attempt to partially answer these questions, Σ was estimated from different simulated datasets and the results were compared to the true values. The datasets were created according to the description in the previous section and the same parameters were used, i.e. $N_G = 10000$, $N_I = 4$ and $\alpha = 2$. Five different cases, listed in Table 1, were examined. As

Case	Correlation	Heavy tails	Regulated genes	Filter
I	No	No	None	No
II	Yes	No	None	No
III	Yes	Yes	None	No
IV	Yes	No	Yes, 10%	No
V	Yes	No	Yes, 10%	Yes, 5% removed.

Table 1: Descriptions of the five different settings used in this simulation study. When correlations are used, they follow the structure in equation (6).

in the previous section, 100 datasets were simulated for each setting and for each such dataset the covariance matrix Σ and the hyperparameter α were estimated according to Section 4. Table 2 summarises the result where the true value of Σ, the mean value of the estimated Σ as well as the standard deviations are listed. It should be noted that in all cases, except for case **III**, α is estimated with high accuracy and precision.

In the first two cases (**I** and **II**), the covariance matrix was estimated without any bias and with low standard deviation showing that the methods are accurate under the model assumptions. In case **III** the normal distribution was substituted against a t-distribution with 5 degrees of freedom, having substantially heavier tails. The estimated Σ seems to be slightly biased toward higher variances and α was estimated to 1.55 instead of 2. This pattern was also seen when the degrees of freedom were increased to 10 and 15 (results not shown). In case **IV** 10% of the genes were set to be regulated and since no differentially expressed genes are assumed, the regulation leads to positive correlations and increased variance estimates. Having 10% of the genes regulated is a rather high number, but not extreme. Therefore, a filter

was applied to minimise the impact of regulated genes on the estimation of the covariance matrix. For each gene g, the filter calculates the minimal absolute value of the fold change, which will be denoted $X_{g,min}$. Removing the top 5% of the genes with highest $X_{g,min}$ gave a much better estimate of Σ, which is included as case **V**. Note that the genes were only removed from the the estimate of Σ^*, i.e. the arbitrarily scaled Σ, and not from the estimates of α and λ. Also note that the number 5% depends on several parameters, such as the total number of regulated genes and the covariance matrix itself. The results of the filtering procedure on real data is presented in the next section.

	True Σ				Mean estimated Σ				Sample standard deviation			
I	0.50	0.00	0.00	0.00	0.50	0.00	-0.00	-0.00	0.01	0.01	0.01	0.01
	0.00	1.00	0.00	0.00	*0.00*	1.01	-0.00	0.00	*0.01*	0.04	0.02	0.01
	0.00	*0.00*	1.50	0.00	*-0.00*	*-0.00*	1.51	-0.00	*0.02*	*0.02*	0.05	0.02
	0.00	*0.00*	*0.00*	2.00	*-0.00*	*0.00*	*-0.00*	2.02	*0.01*	*0.01*	*0.01*	0.07
II	0.50	0.28	0.17	0.00	0.50	0.28	0.17	0.00	0.02	0.01	0.01	0.01
	0.40	1.00	0.49	0.28	*0.40*	1.00	0.50	0.29	*0.01*	0.04	0.02	0.03
	0.20	*0.40*	1.50	0.69	*0.20*	*0.40*	1.51	0.70	*0.01*	*0.01*	0.06	0.04
	0.00	*0.20*	*0.40*	2.00	*0.00*	*0.20*	*0.40*	2.00	*0.01*	*0.01*	*0.01*	0.11
III	0.50	0.28	0.17	0.00	0.51	0.29	0.18	-0.00	0.02	0.01	0.01	0.01
	0.40	1.00	0.49	0.28	*0.40*	1.01	0.50	0.28	*0.01*	0.04	0.02	0.02
	0.20	*0.40*	1.50	0.69	*0.20*	*0.40*	1.52	0.70	*0.01*	*0.01*	0.05	0.03
	0.00	*0.20*	*0.40*	2.00	*-0.00*	*0.20*	*0.40*	2.03	*0.01*	*0.01*	*0.01*	0.07
IV	0.50	0.28	0.17	0.00	0.61	0.39	0.28	0.11	0.02	0.02	0.02	0.01
	0.40	1.00	0.49	0.28	*0.48*	1.11	0.60	0.39	*0.01*	0.04	0.03	0.01
	0.20	*0.40*	1.50	0.69	*0.28*	*0.45*	1.61	0.80	*0.01*	*0.01*	0.06	0.04
	0.00	*0.20*	*0.40*	2.00	*0.10*	*0.25*	*0.43*	2.11	*0.01*	*0.01*	*0.01*	0.08
V	0.50	0.28	0.17	0.00	0.46	0.21	0.11	-0.02	0.01	0.01	0.01	0.02
	0.40	1.00	0.49	0.28	*0.33*	0.90	0.38	0.22	*0.01*	0.02	0.02	0.02
	0.20	*0.40*	1.50	0.69	*0.14*	*0.34*	1.39	0.59	*0.01*	*0.02*	0.06	0.03
	0.00	*0.20*	*0.40*	2.00	*-0.02*	*0.16*	*0.36*	1.93	*0.02*	*0.01*	*0.01*	0.07

Table 2: Result from the estimations of Σ from each of the five different cases. Correlations are shown in italic and covariances in non-italic. The parameter values used were $N_G = 10000$, $N_I = 4$ and $\alpha = 2$. The mean values and sample standard deviations were calculated from the result of 100 simulated dataset. Refer to Table 1 for a description of the different cases.

6 Results from real data

WAME was run on three real data sets: the ischemic part of the dataset of Hall et al. (2004), the dataset of Benson et al. (2004) (henceforth referred to as the *Cardiac* and *Polyp* datasets, respectively) and the *Swirl* dataset (described in Section 3.3 of Dudoit and Yang, 2003). These datasets represent microarray experiments with different characteristics; different laboratories, both two-colour cDNA and one-channel oligonucleotide (Affymetrix) arrays, different tissues and two different species (human and zebrafish). The Cardiac and Swirl datasets are publicly available.

The Cardiac dataset is described to have been strictly quality controlled by a combination of several available methods. The dataset is therefore interesting to examine to see if WAME detects relevant differences in quality even in an example of a quality controlled, publicly available dataset. The Polyp dataset includes one biopsy that was previously thought to be an outlier and therefore discarded, thus providing a case with one seemingly lesser quality to be detected. In the Swirl dataset, two highly differentially expressed genes exist. Therefore, it is of interest to check that those genes are highly ranked by WAME. Furthermore, the Swirl dataset has been analysed in e.g. (Smyth, 2004).

6.1 Cardiac dataset

In the public dataset from Hall et al. (2004), heart biopsies from 19 patients with heart failure were harvested before and after mechanical support with a ventricular assist device. The aim of the study was to "define critical regulatory genes governing myocardial remodelling in response to significant reductions in wall stress", where a first step was to identify differentially expressed genes between the two conditions.

Affymetrix one-channel oligonucleotide arrays of type HG-U133A were used in the study, each containing 22283 probe-sets. The quality of the arrays was controlled using quality measures recommended by Affymetrix as well as by the program Gene Expressionist (GeneData, Basel, Switzerland). The quality of the different lab steps leading to the actual hybridisations were controlled using standard methods. The 19 patients were divided into three groups: ischemic (5 patients), acute myocardial infarction (6 patients) and non-ischemic (8 patients). The ischemic group was the smallest and consequently the one where quality variations might make the biggest difference. It was therefore chosen for further examination using WAME, to see if relevant quality variations could be detected despite the close quality

monitoring.

The dataset was retrieved in raw .CEL-format from the public repository Gene Expression Omnibus (Edgar et al., 2002). The .CEL-files were subsequently processed using RMA (Irizarry et al., 2003) on all the arrays of the 19 patients simultaneously. Patient-wise $\log_2$-ratios of the five ischemic patients were then formed by taking pairwise differences of the $\log_2$ measurements before and after implant.

Applying WAME to the patient-wise $\log_2$-ratios provided interesting results. The estimated covariance matrix (see Table 3) suggests that two of the five patients (I13 and I7) were substantially more variable than the others, while the correlations between patients were rather limited. These numbers seem credible when examining Figure 3, where for each pair of patients, the respective $\log_2$-ratios of all genes were plotted against each other. The plots clearly imply that the observations of the two patients in question (I13 and I7) are more variable than the others.

The corresponding weights, derived from the estimated covariance matrix Σ, are shown in Table 4. As was discussed in Sections 4.1 and 5.2, when estimating Σ all genes are assumed to be non-differentially expressed. To examine the impact of potentially regulated genes on the estimation of Σ, the analysis was redone, removing genes with high lowest absolute $\log_2$-ratio in the estimation of Σ, as described in Section 5.2. The individual elements of the estimated covariance matrix and of α changed only slightly, even when as much as 50% of the data was removed (data not shown). This is reflected in the weights in Table 4.

	Patient				
Patient	I12	I13	I4	I7	I8
I12	0.046	0.003	0.001	0.012	0.002
I13	*0.033*	0.196	-0.014	0.007	-0.001
I4	*0.023*	*-0.126*	0.065	0.013	0.002
I7	*0.111*	*0.030*	*0.102*	0.258	-0.017
I8	*0.040*	*-0.011*	*0.038*	*-0.152*	0.047

Table 3: Estimated covariance-correlation matrix, Σ, for patients in the Cardiac dataset. (Correlations in italic, covariances in non-italic.)

The hyperparameter α related to the spread of the gene-specific variance scaling factors, c_g, was estimated to 1.92, giving a thick tail for the prior distribution. Thus removing c_g by transformation when estimating Σ (Section 4.1) is justified.

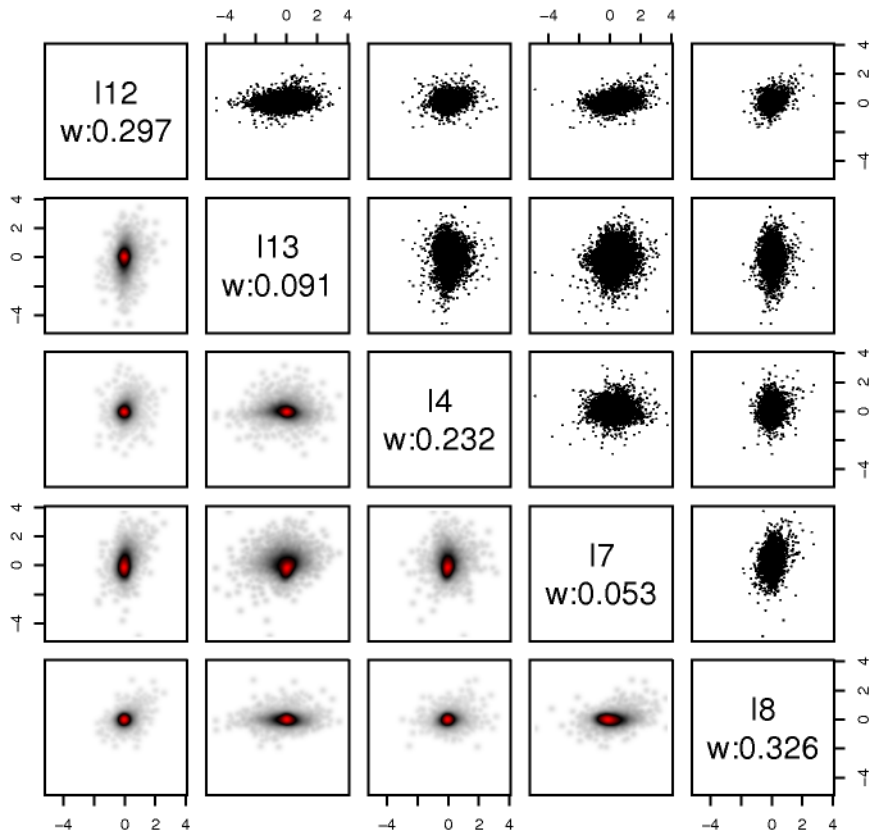


Figure 3: Pair-wise plots of the $\log_2$-ratios of the patients in the Cardiac dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots are solid black (cf. Figure 6 in Huber et al., 2003). The colour-scale is, in increasing level of density: white, grey, black and red.

	Patient				
Removed genes	I12	I13	I4	I7	I8
none	0.297	0.091	0.232	0.053	0.326
5%	0.301	0.089	0.233	0.054	0.323
10%	0.303	0.087	0.235	0.053	0.321
50%	0.323	0.082	0.240	0.046	0.308

Table 4: Weights for patients in the Cardiac dataset. Different numbers of potentially regulated genes were removed in the estimation of Σ, to check their influence. Potential regulation was measured by minimal absolute $\log_2$-ratio among the patients.

Inspecting the fitted distribution of S_g given $\alpha = 1.92$ against the empirical distribution of S_g reveals a good fit (see Figure 4), implying that the family of inverse gamma prior distributions is rich enough for this dataset.

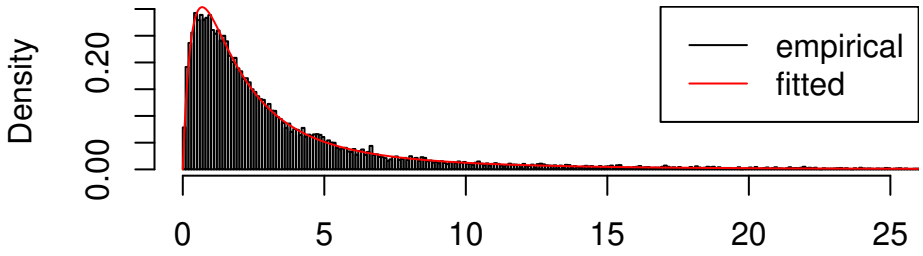


Figure 4: Empirical distribution of S_g in the Cardiac dataset, together with the density of S_g given $\alpha = 1.92$.

Examining the observed values of the statistic, T_g, compared to the expected null distribution reveals a good overall concordance (see Figure 5). Some genes have a larger t_g than can be explained by the null distribution, which points toward some of them being up-regulated by the treatment (see the qq-plot in Figure 5).

6.2 Polyp dataset

In the dataset from Benson et al. (2004), biopsies from nasal polyps of five patients were taken before and after treatment with local glucocorticoids. The goal was to examine closer the mechanisms behind the effect of the

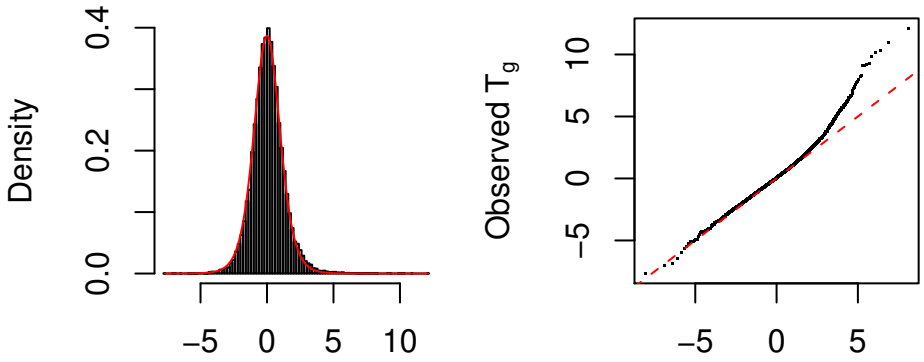


Figure 5: To the left, a histogram of the observed T_g-values together with the density of the null distribution (in red), in the Cardiac dataset. To the right, a quantile-quantile plot where the observed values of T_g are paired with the quantiles of T_g under the null hypothesis. The central part of the empirical distribution follows the identity line well, showing good concordance with the null distribution. For high positive T_g-values, the observations clearly deviate from the predicted ones, pointing at the existence of up-regulated genes.

treatment and one step was to identify differentially expressed genes. Technical duplicates stemming from the same extracted RNA were run for each biopsy on Affymetrix HG-U133A arrays, forming a dataset of 20 arrays and 22283 probe-sets.

Comparing each of the arrays in the dataset with all arrays from other patients and/or conditions, by looking at pair-wise scatterplots, the arrays from before treatment of patient 2 consistently showed larger variation than any other. The biopsy in question was found to be considerably smaller than the others, providing possible explanations by e.g. non-representativeness in tissue distribution. The data from patient 2 was therefore excluded in Benson et al. (2004).

WAME would preferably identify the patient 2 observation as having larger variation and downweight it. The data was processed using RMA (Irizarry et al., 2003) and the $\log_2$-ratio for each patient was formed by taking differences between the averages over the technical duplicates, before and after treatment, combining 4 arrays for each patient into one set of $\log_2$-ratios. Making one scatter plot of the two sets of $\log_2$-ratios for each pair of patients (Figure 6) clearly indicates that patient 2 is more variable than

patients 1,3 and 5. Interestingly, the measurements from patients 1 and 2 seem to be highly correlated and patient 4 seems to have high variability.

Estimating the covariance matrix, Σ, the correlation between patients 1 and 2 is estimated to 0.82 (see Table 5), which is high but not unbelievable when studying Figure 6. The variance of patient 2 is furthermore estimated to four times that of patient 1. Examining the resulting weights, patient 2 actually receives a weight of -2% (see Table 6). The negativeness is a result of it's variance being much higher than that of patient 1, together with them being highly correlated. As negative weights seem questionable, a natural solution is to remove patient 2, which was done in (Benson et al., 2004). Beside the result of the very low weight for patient 2, the other patients receive distinctly different weights, which is interesting.

	Patient				
Patient	1	2	3	4	5
1	0.300	0.493	0.000	-0.012	-0.067
2	*0.822*	1.200	0.004	0.041	-0.157
3	*0.002*	*0.012*	0.091	-0.071	-0.055
4	*-0.038*	*0.067*	*-0.417*	0.319	0.102
5	*-0.291*	*-0.340*	*-0.434*	*0.430*	0.178

Table 5: Estimated covariance-correlation matrix, Σ, for patients in the Polyp dataset. (Correlations in italic, covariances in non-italic.)

	Patient				
Removed genes	1	2	3	4	5
none	0.179	-0.026	0.483	0.104	0.260
5%	0.181	-0.025	0.481	0.104	0.259
10%	0.180	-0.024	0.482	0.103	0.259
50%	0.157	-0.015	0.506	0.100	0.252

Table 6: Weights for the patients in the Polyp dataset. Different numbers of potentially regulated genes were removed, to check their potential influence in the estimation of Σ. Potential regulation was measured by minimal absolute $\log_2$-ratio among the patients.

The hyperparameter α, related to the spread of the gene-specific variance scaling factors, c_g, was estimated to 1.97, giving infinite variance for the

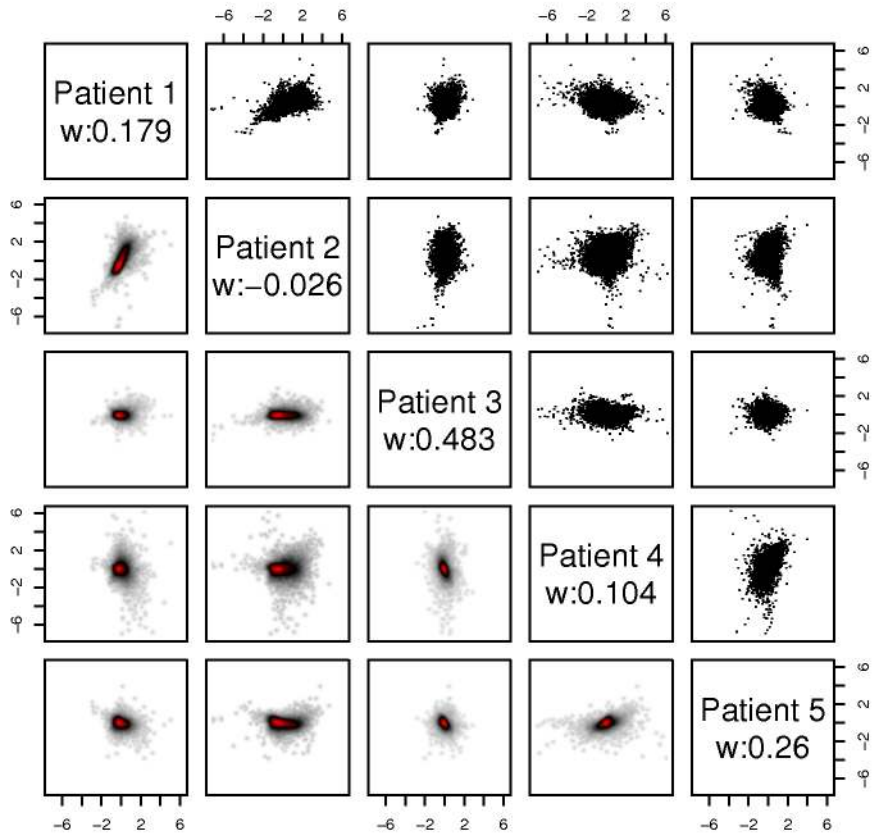


Figure 6: Pair-wise plots of the $\log_2$-ratios of the patients in the Polyp dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots are solid black (cf. Figure 6 in Huber et al., 2003). The colour-scale is, in increasing level of density: white, grey, black and red.

distribution of c_g. The fit of S_g given $\alpha = 1.97$ was very good (see Figure 10 in the Appendix).

As in the Cardiac dataset, the weights were steadily estimated when potentially regulated genes were removed in the estimation of the covariance matrix Σ (see Table 6). The estimated correlations between patients 3, 4 and 5 were reduced somewhat. Removing 5% of the genes reduced those correlations by 0.03-0.04 and removing 10% reduced them by 0.06-0.07. The high correlation between patient 1 and 2 was only slightly reduced (<0.03), even when 50% of the genes were removed.

Examining the observed values of the statistic, T_g, compared to the expected null distribution (see Figure 7) reveals a good overall concordance. Some genes have a more extreme T_g than can be explained by the null distribution, which points toward many of them being regulated by the treatment (see the qq-plot in Figure 7).

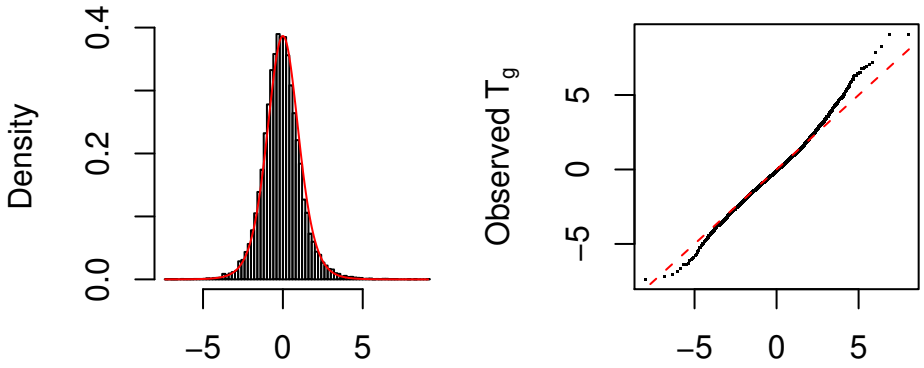


Figure 7: To the left, a histogram of the observed T_g-values together with the density of the null distribution (in red), in the Polyp dataset. To the right, a quantile-quantile plot where the observed values of T_g are paired with the quantiles of T_g under the null hypothesis. The central part of the empirical distribution follows the identity line well, showing good concordance with the null distribution. For extreme T_g-values, the observations clearly deviate from the predicted ones, pointing at the existence of regulated genes.

6.3 Swirl dataset

In the Swirl experiment (described on page 80 in Dudoit and Yang, 2003), one goal was to identify genes that are differentially expressed in zebrafish

carrying a point mutated SRB2 gene, compared to ordinary, wild-type zebrafish. SRB2 and one of it's known targets, Dlx3 are expected to be highly differentially expressed in this experiment, thus these genes should be highly ranked using WAME. The Swirl dataset has been examined in Smyth (2004).

The dataset consists of four two-colour cDNA microarrays with 8448 spots, with publicly available data. We used standard pre-processing to compensate for effects such as background and dye bias (background correction *subtract* and within-array normalisation *print tip loess* were used in the LIMMA package (Smyth et al., 2003)). Between-array scale normalisation (Yang et al., 2002) was not performed in contrast to the analysis in Smyth (2004). When including between-array scale normalisation in combination with LIMMA in the simulation study of Section 5.1 the performance was not notably increased (results not shown).

Making one scatter plot of the $\log_2$-ratios for each pair of arrays (Figure 8) indicates that array 2 is less variable than the others, while the genes with lowest $\log_2$-ratio on array 1 seem to be outliers, since they are not extreme in any other array. Examining the estimated covariance matrix (see Table 7), array 2 indeed receives the highest variance. In addition, there are substantial correlations between arrays 1-3, 2-4 and 3-4, which is also indicated by the scatter-plots (Figure 8).

	Array			
Array	1	2	3	4
1	0.128	0.007	0.079	0.017
2	*0.066*	0.086	-0.002	0.038
3	*0.489*	*-0.017*	0.203	0.076
4	*0.136*	*0.371*	*0.482*	0.124

Table 7: Estimated covariance-correlation matrix, Σ, for the arrays in the Swirl dataset. (Correlations in italic, covariances in non-italic.)

When re-performing the estimation of Σ , removing potentially regulated genes (in analogy with the analyses of the Polyp and Cardiac datasets), the correlations were decreased somewhat. Removing 5% of the genes decreased the three high correlations by 0.02-0.06, while removing 10% decreased them by 0.04-0.08. However, the corresponding weights only changed marginally (see Table 8).

The hyperparameter α was estimated to 1.89. Further analysis of the dataset shows that the distribution of S_g fits the predicted distribution of S_g

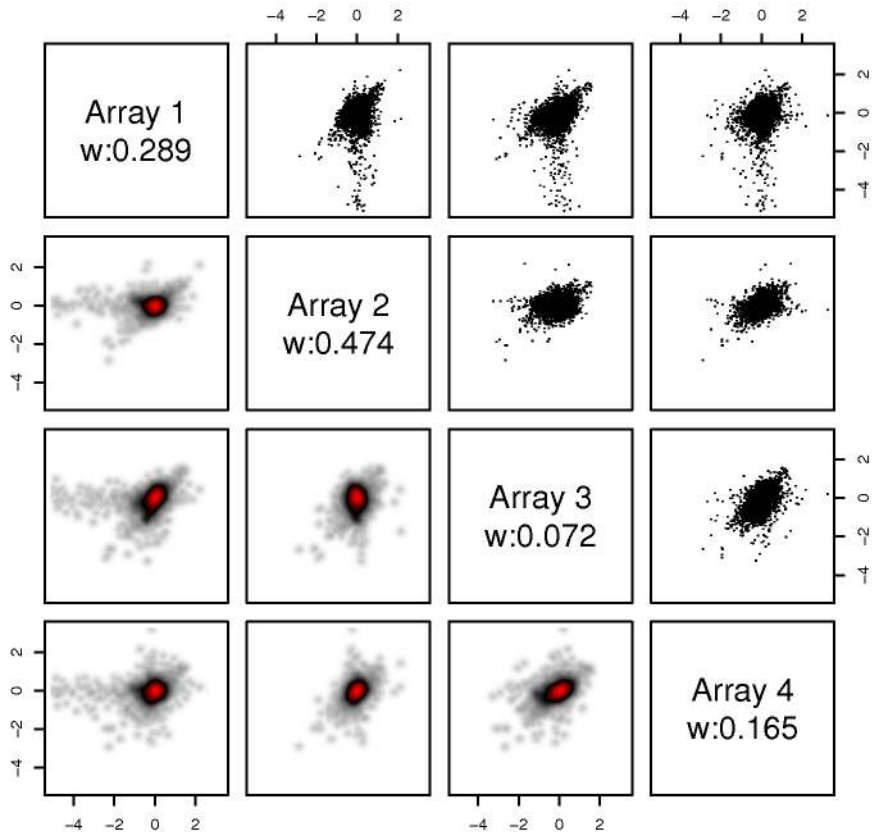


Figure 8: Pair-wise plots of the $\log_2$-ratios of the arrays in the Swirl dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots are solid black (cf. Figure 6 in Huber et al., 2003). The colour-scale is, in increasing level of density: white, grey, black and red.

	Array			
Removed genes	1	2	3	4
none	0.289	0.474	0.072	0.165
5%	0.288	0.469	0.076	0.166
10%	0.290	0.462	0.075	0.173
50%	0.282	0.447	0.087	0.184

Table 8: Weights for the arrays in the Swirl dataset. Different numbers of potentially regulated genes were removed, to check their potential influence in the estimation of Σ. Potential regulation was measured by minimal absolute $\log_2$-ratio among the arrays.

given $\alpha = 1.89$ well (see Figure 11 in in the Appendix). The observed values of the statistic, T_g, seem to fit the null distribution well (see Figure 9).

Since the point mutated gene, SRB2 and one of it's known targets, Dlx3, are expected to be highly differentially expressed, their actual ranking is of interest. In Table 9 below, the top 20 genes as ranked by WAME are listed. The values of some widely used statistics are included for comparison. The rankings by WAME and the moderated t-statistic (Smyth et al., 2003) are quite similar, while the rankings by the ordinary t-statistic and the average $\log_2$-ratio (i.e. fold change) are rather different than the one by WAME, which was expected. All four spots for the two validated genes are included in WAME:s top 20 list (see Table 9).

7 Discussion

A drawback of the microarray technology is that it involves several consecutive steps, each exhibiting large quality variation. Thus there is a strong need for quality assessment and quality control to handle occurrences of poor quality. In this paper, we introduce a method called WAME for the analysis of paired microarray experiments, which aims at estimating array-wide quality deviations and integrates these quality estimates into the statistical analysis.

The quality deviations are modelled as different variances for different repetitions (e.g. arrays) as well as correlations between them in a covariance matrix σ, thus catching both unequal precision and systematic errors. These are contained in a covariance matrix Σ. Genes have different variability (both biological and technical), which is modelled by a gene-specific variance

Name	ID	average $\log_2$-ratio	ordinary t-statistic	moderated t-statistic	WAME
fb85d05	18-F10	-2.66	-18.41	-20.79	-15.15
fb58g10	11-L19	-1.60	-14.32	-14.15	-11.51
control	Dlx3	-2.19	-15.91	-17.57	-11.17
control	Dlx3	-2.19	-13.58	-16.08	-9.84
fb24g06	3-D11	1.32	19.52	13.62	9.80
fb54e03	10-K5	-1.20	-25.74	-13.11	-9.66
fc22a09	27-E17	1.26	24.76	13.68	9.50
fb40h07	7-D14	1.35	14.15	12.69	9.12
fb85a01	18-E1	-1.29	-17.35	-13.01	-8.81
fb87f03	18-O6	-1.08	-27.90	-12.06	-8.80
fb37e11	6-G21	1.23	14.37	11.94	8.47
fb94h06	20-L12	1.28	15.41	12.54	8.46
fb87d12	18-N24	1.28	12.96	11.87	8.39
control	BMP2	-2.24	-8.63	-11.78	-8.33
fc10h09	24-H18	1.20	15.05	11.92	8.23
fb85f09	18-G18	1.29	11.50	11.38	8.22
control	BMP2	-2.33	-8.37	-11.58	-7.95
fb26b10	3-I20	1.09	15.50	11.17	7.81
fb37b09	6-E18	1.31	11.57	11.55	7.78
fc22f05	27-G10	-1.19	-10.42	-10.44	-7.70

Table 9: The top 20 most probably regulated genes in the Swirl dataset according to WAME.

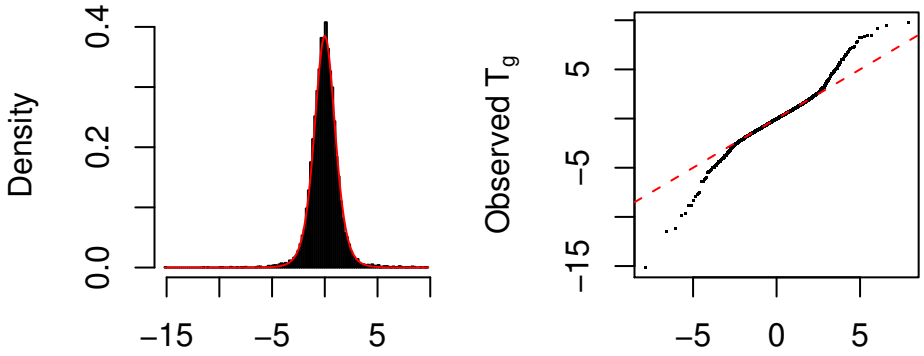


Figure 9: To the left, a histogram of the observed T_g-values together with the density of the null distribution (in red), in the Swirl dataset. To the right, a quantile-quantile plot where the observed values of T_g are paired with the quantiles of T_g under the null hypothesis. The central part of the empirical distribution follows the identity line well, showing good concordance with the null distribution. For extreme T_g-values, the observations clearly deviate from the predicted ones, pointing at the existence of regulated genes.

scaling factor c_g. Given this structure, the pair-wise measured $\log_2$-ratios for each gene are assumed to be normally distributed. It should be straightforward to incorporate exclusion of outlying gene-specific observations (e.g. spots) into the model. Including quantitative measures of quality of such observations, e.g. by a hierarchical variance component model (cf. Bakewell and Wit (2005)), would be interesting as future work.

Estimation of the covariance matrix is non-trivial due to the gene-specific scaling factors and unknown differential expressions μ_g. Here, an assumption is made that most genes are not differentially expressed ($\mu_g = 0$) and a transformation is performed to remove the gene-specific scaling factors. Then, a scaled version of Σ is estimated using numerical maximum likelihood, based on the derived resulting distribution. The assumption of no differential expression somewhat limits the experimental setups that can be analysed. However, this is not as consequential as it might sound, since it is made by most of the procedures that have become *de facto* standard in the (preceding) normalisation step.

Since most microarray experiments contain only a few repetitions, the estimate of the gene-specific variance scaling factor c_g is imprecise, which can easily lead to false conclusions if not accounted for. Here, an empirical Bayes

approach is taken where an inverse gamma prior distribution is assumed, in effect moderating extreme estimates (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Smyth, 2004). The hyperparameter α determining the spread of the prior distribution is estimated from the data, by numerical maximum likelihood together with the scale of the previously estimated arbitrarily scaled Σ.

To identify differentially expressed genes a likelihood-ratio test is derived, resulting in the *weighted moderated t-statistic*, which is a generalisation of the moderated t-statistic in Smyth (2004). Here, the estimated covariance matrix Σ is used both to produce weights for the different repetitions and gene-specific variance estimates. The weighted mean is the estimate of differential expression with minimal variance.

As discussed above, array-wide quality deviations in all steps leading to the observed $\log_2$-ratios are estimated and incorporated into the analysis. The current paper is restricted to paired two-sample settings where most genes are non-differentially expressed. A generalisation similar to (Smyth, 2004) should be possible to make, for experiments restricted to pairwise measurements with most genes being non-differentially expressed. The scaled estimate of the covariance matrix Σ could be calculated according to the procedure in the current paper (cf. Section 4.1). The unknown scale of the covariance estimate, as well as the parameter α of the prior distribution for the gene-specific variance scales, could be estimated utilising generalised residual sums of squares for all genes, appropriately defined through the norm determined by Σ (cf. S_g in Section 4.2). Tests for single or multiple identifiable linear combinations of the featuring expected values could then be derived similar to in the current paper, forming weighted moderated t-statistics and modified F-statistics. Work on a generalisation, with simulated and real data sets is in progress.

A simulation study was done to compare the performance of WAME to four published methods. On data without correlations and with equal variances between repetitions, WAME performs as well as the moderated t-statistic which assumes this structure. When correlations and/or unequal variances were included, WAME performs better than all the other methods. In one case, using WAME results in almost a third less false positives which can correspond to hundreds of genes. Evaluating the point estimation of the covariance matrix Σ revealed good precision and accuracy when no regulated genes were present. Including 10% regulated genes resulted in a bias, which was partly handled by removing genes likely to be regulated. In both cases estimation of the hyperparameter α was nearly unbiased and accurate. The estimate of Σ was essentially unbiased when heavy tails was introduced in

contrast to the estimate of α which was estimated to 1.55 instead of 2.

Three real datasets were analysed: the ischemic part of the dataset of Hall et al. (2004)(publicly available), the dataset of Benson et al. (2004) and the *Swirl* dataset (described in chapter 3.3 of Dudoit and Yang, 2003)(publicly available). In all cases, relevant correlations and differences in precision between replicates were found, even in first dataset which had been quality controlled by a combination of several available methods. The exact origin of the correlations is an interesting, open question. In the second dataset one previously identified outlier was practically removed by WAME. In the Swirl dataset, expected differentially expressed genes are ranked among the top 20. Relevant empirical distributions showed good fit to the theoretic distributions, pointing toward the family of prior distributions for c_g being flexible enough and the normal assumption being satisfactory.

The model used in WAME is optimistic in many ways. For example, exact normality is not to be expected and the independence between the genes is hard to fully motivate. The noise structure might also be different for the regulated genes, e.g. if there are several normalising procedures involved in the pre-processing step. This would certainly affect the power and possibly point towards the rationality of using a moderated impact of Σ on the weights in the final analysis. Thus, even if simulations under the model assumptions show very promising results, there are many experimental situations where the model assumptions and thus the theoretical performance are not fully justified. We intend to look further into different robustness questions for model deviations in the future.

It is also important keep the main role of microarrays in mind, in which tests of regulation of tens of thousands of genes is an exploratory tool for deriving candidate ranking lists of potentially regulated genes, that in the next steps will be biologically interpreted and validated by more precise techniques. We claim that our approach competes well with other methods in the production of such lists.

To summarise, WAME estimates and integrates array-wide quality deviations into the analysis of paired microarray experiments. An empirical Bayes approach is used to moderate the gene-specific variance scale estimates, resulting in a weighted moderated t-statistic with a derived distribution. The performance of WAME has been evaluated on both simulated and real microarray data, with interesting results.

Acknowledgements

We would like to thank Mikael Benson, Lars Olaf Cardell, Lena Carlsson and Margareta Jernås for valuable discussions and access to the data from (Benson et al., 2004).

Appendix

Additional Figures

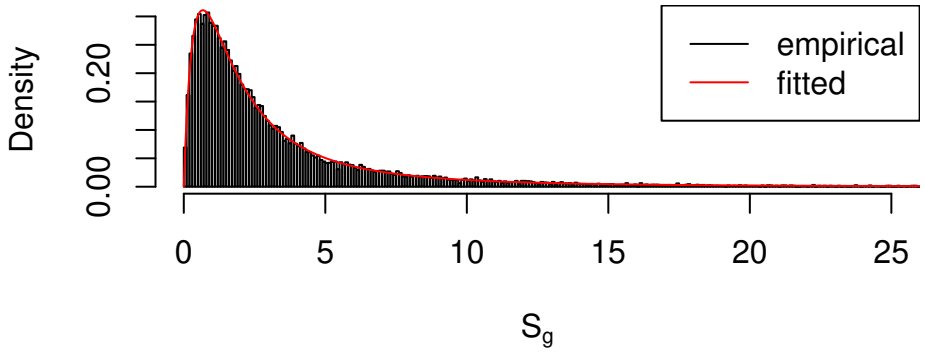


Figure 10: Empirical distribution of S_g in the Polyp dataset, together with the density of S_g given $\alpha = 1.97$.

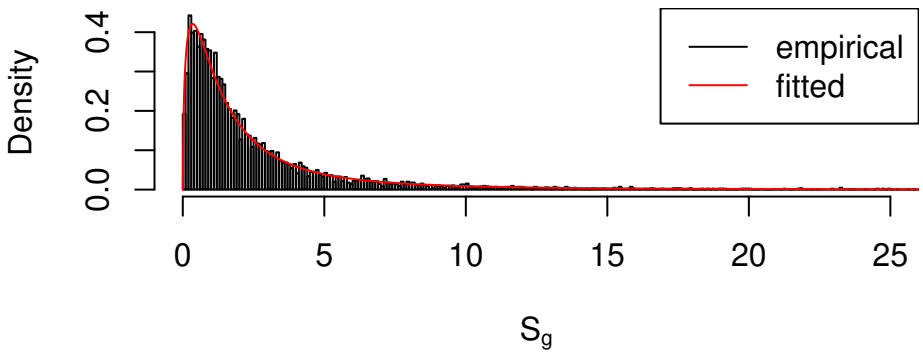


Figure 11: Empirical distribution of S_g in the Swirl dataset, together with the density of S_g given $\alpha = 1.89$.

Mathematical details

We observe $\mathbf{X}_g = (X_{g1}, \ldots, X_{gN_I})$ where $g = 1, \ldots, N_G$. Let Σ be a covariance structure matrix for the N_I repetitions, c_g a set of gene-specific variance scaling factors and α a hyperparameter determining the shape of the prior distribution for c_g. Then for fixed μ_g, Σ and α,

$$\begin{aligned} c_g &\sim \Gamma^{-1}(\alpha, 1), \text{ and} \\ \mathbf{X}_g \mid c_g &\sim \mathbf{N}_{N_I}(\mu_g \mathbf{1}, c_g \Sigma) \end{aligned}$$

and all variables corresponding to different genes are assumed independent.

Estimation of a scaled version of the matrix Σ

Assume that $\mu_g = 0$ for all g. Under this assumption, it is possible to derive a scale independent estimate of the covariance matrix Σ by a transformation of the vector $\mathbf{X}_g$. This is done as follows (the index g is dropped to increase the readability) Let $\mathbf{U} = (U_1, \ldots, U_{N_I})$ where

$$U_i = \begin{cases} X_1 & \text{if } i = 1 \\ X_i / X_1 & \text{if } 2 \leq i \leq N_I. \end{cases}$$

The inverse becomes

$$X_i = \begin{cases} U_1 & \text{if } i = 1 \\ U_i U_1 & \text{if } 2 \leq i \leq N_I. \end{cases}$$

and the Jacobian it

$$J(u_1, \ldots, u_{N_I}) = u_1^{N_I - 1},$$

so for $\mathbf{U} \in \mathbb{R}^{N_I}$ the density becomes

$$\begin{aligned} f_{\mathbf{U} \mid c,\Sigma}(\mathbf{u}) &= f_{\mathbf{X} \mid c,\Sigma}\left(\mathbf{x}(\mathbf{u})\right) |J(\mathbf{u})| \\ &= (2\pi)^{-N_I/2} \, c^{-N_I/2} \, |\Sigma|^{-1/2} \, |u_1|^{N_I - 1} e^{-\frac{u_1^2}{2c} v^{\mathrm{T}} \Sigma^{-1} v}. \end{aligned}$$

where $v = (1, u_2, \ldots, u_{N_I})^{\mathrm{T}}$. Integration over u_1 yields

$$\begin{aligned} f_{U_2,\ldots,U_{N_I} \mid \Sigma}(u_2, \ldots, u_{N_I} \mid \Sigma) &= \int_{-\infty}^{\infty} f_{\mathbf{U} \mid c,\Sigma}(\mathbf{u} \mid c, \Sigma) \, du_1 \\ &= C \, |\Sigma|^{-1/2} \left[v^{\mathrm{T}} \Sigma^{-1} v \right]^{-N_I/2}, \end{aligned} \tag{7}$$

where C is a normalisation constant and v is defined as above. This density is scale invariant with respect to the parameter Σ in the sense that for any scalar λ,

$$f_{U_2,\ldots,U_{N_I} \mid \Sigma}(u_2, \ldots, u_{N_I} | \lambda\Sigma) = f_{U_2,\ldots,U_{N_I} \mid \Sigma}(u_2, \ldots, u_{N_I} | \Sigma).$$

Thus, it is also independent of c and under the assumption of independent genes, the log-likelihood function becomes

$$l(\Sigma) = C' - \frac{N_G}{2} \log\left(|\Sigma|\right) - \frac{N_I}{2} \sum_{g=1}^{N_g} \log\left(v_g^{\mathrm{T}} \Sigma^{-1} v_g\right),$$

where C' is a constant that is independent of Σ. Numerical maximisation yields a scaled version of Σ, denoted Σ^*. Here the first element in the first row of Σ^* is fixed to one.

Estimation of the hyperparameter α and the scale λ

From the model assumptions, we know that

$$c_g \sim \Gamma^{-1}(\alpha, 1).$$

Assume that Σ is known and define

$$S_g = (A\mathbf{X}_g)^{\mathrm{T}} (A\Sigma A^{\mathrm{T}})^{-1} A\mathbf{X}_g,$$

where A is a contrast matrix, i.e. a matrix of dimension $N_I - 1 \times N_I$, with full rank and with each row sum equal to 0. It follows that

$$S_g \sim c_g \times \chi^2_{N_I - 1}.$$

The unconditional distribution of S_g can be derived by integrating over c_g, i.e.,

$$\begin{aligned} f_{S_g \mid \alpha}(s_g) &= \int_0^\infty f_{S_g \mid c_g}(s) f_{c_g \mid \alpha}(c_g)\, dc_g \\ &= \frac{1}{2} \frac{(s/2)^{(N_I-1)/2-1}}{\Gamma(\alpha)\,\Gamma((N_I-1)/2)} \int_0^\infty c^{-\alpha-(N_I-1)/2-1} e^{-(s/2+1)}\, dc_g \\ &= \frac{1}{2} \frac{\Gamma(\alpha+(N_I-1)/2)}{\Gamma(\alpha)\Gamma((N_I-1)/2)} \frac{(s/2)^{(N_I-1)/2-1}}{[1+s/2]^{\alpha+(N_I-1)/2}}. \end{aligned}$$

This is a beta prime distribution (also called a beta distribution of the second kind) (Johnson et al., 1995) with parameters $N_I - 1$ and α which is denoted $\beta'(N_I-1, \alpha)$. Since only a scaled version of Σ, denoted Σ^*, is assumed known from the primary estimation step, the following entities are defined. Let

$$\begin{aligned} \Sigma^* &= \lambda\Sigma \\ S_g^* &= (A\mathbf{X}_g)^{\mathrm{T}}(A\Sigma^* A^{\mathrm{T}})^{-1} A\mathbf{X}_G = S_g/\lambda, \end{aligned}$$

where λ in the unknown scale for Σ^*. It follows that

$$S_g^* \sim 2/\lambda \times \beta'(N_I - 1, \alpha).$$

The log likelihood function can be simplified to

$$\begin{aligned} l(\alpha, \lambda | \{s_g\}_{g=1}^{N_G}) = C + N_G \left[(N_I-1)/2 \log(\lambda) + \log\Gamma(\alpha + (N_I-1)/2) - \log\Gamma(\alpha)\right] \\ - (\alpha + (N_I-1)/2) \sum_{g=1}^{N_G} \log(s_g\lambda/2 + 1). \end{aligned}$$

Numerical maximum likelihood is used to estimate α and λ, which together with Σ^* can be used to calculate an estimate for Σ.

Inference about μ_g

The hypotheses that are interesting to test are if different genes are regulated or not, that is for each g,

$$\begin{aligned} H_0 &: \text{ gene } g \text{ is not regulated } (\mu_g = 0) \\ H_A &: \text{ gene } g \text{ is regulated } (\mu_g \neq 0). \end{aligned}$$

To test these hypotheses a maximum likelihood ratio (LRT) test is derived. For each g, we reject H_0 if

$$\frac{\sup_{\mu_g \neq 0} L(\mu_g | \mathbf{x}_g)}{L(0|\mathbf{x}_g)} \geq k,$$

where $1 \leq k < \infty$. The likelihood L can be calculated by integration over c_g, i.e.

$$L\left(\mu_g|\mathbf{x}\right) = \int f_{\mathbf{X}\,|\,\mu_g,c_g,\Sigma}(\mathbf{x}) f_{c_g\,|\,\alpha}(c_g)\; dc_g$$

$$= (2\pi)^{-N_I/2} \left|\Sigma\right|^{-1/2} \frac{\Gamma(N_I/2+\alpha)}{\Gamma(\alpha)} \left[\frac{\left(\mathbf{x}_g - \mu_g\mathbf{1}\right)^{\mathrm{T}} \Sigma^{-1} \left(\mathbf{x}_g - \mu_g\mathbf{1}\right)}{2} + 1\right]^{-N_I/2-\alpha}.$$

To calculate the numerator in the likelihood ratio we need to maximise L over μ_g, which is the same as minimising

$$\left(\mathbf{x}_g - \mu_g\mathbf{1}\right)^{\mathrm{T}} \Sigma^{-1} \left(\mathbf{x}_g - \mu_g\mathbf{1}\right).$$

A little algebra shows that this optimum corresponds to the argument

$$\hat{\mu}_g = \frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\mathbf{x}_g\,.$$

We will use $\bar{x}_g^w$ to denote this weighted average and it can be shown to be the weighted mean with least variance. The maximum value of the likelihood function becomes

$$L(\bar{x}_g^w|\mathbf{x}_g) = (2\pi)^{-N_I/2} \left|\Sigma\right|^{-1/2} \frac{\Gamma(N_I/2+\alpha)}{\Gamma(\alpha)} \left[\frac{\mathbf{x}_g^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_g - (\bar{x}_g^w)^2\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}{2} + 1\right].$$

Using this, the likelihood ratio test statistic can be rewritten as

$$\begin{aligned}
\frac{L\left(\bar{x}_g^w|\mathbf{x}_g\right)}{L\left(0|\mathbf{x}_g\right)} &= \left[\frac{\mathbf{x}_g^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_g + 2}{\mathbf{x}_g^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_g - \left(\bar{x}_g^w\right)^2\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1} + 2}\right]^{N_I/2+\alpha} \\
&= \left[1 + \frac{\left(\bar{x}_g^w\right)^2\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}{\mathbf{x}_g\Sigma^{-1}\mathbf{x}_g - \left(\bar{x}_g^w\right)^2\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1} + 2}\right]^{N_I/2+\alpha} \\
&= \left[1 + \frac{\left(\bar{x}_g^w\right)^2\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}{\left(\mathbf{x}_g - \left(\bar{x}_g^w\right)\mathbf{1}\right)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{x}_g - \left(\bar{x}_g^w\right)\mathbf{1}\right) + 2}\right]^{N_I/2+\alpha} \\
&= \left[1 + \frac{(\bar{x}_g^w)^2\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}{\left(A_{\mathbf{w}}\mathbf{x}_g\right)^{\mathrm{T}}\Sigma^{-1}\left(A_{\mathbf{w}}\mathbf{x}_g\right) + 2}\right]^{N_I/2+\alpha}
\end{aligned}$$

where $A_{\mathbf{w}}$ is the contrast matrix

$$A_{\mathbf{w}} = \begin{pmatrix}
1-w_1 & -w_2 & -w_3 & \dots & -w_{N_I} \\
-w_1 & 1-w_2 & -w_3 & \dots & -w_{N_I} \\
\dots & \dots & \dots & \dots & \dots \\
-w_1 & -w_2 & -w_3 & \dots & 1-w_{N_I}
\end{pmatrix}$$

and w_i is the i:th element of the vector

$$\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}.$$

The next step is to show that

$$(A_{\mathbf{w}}\mathbf{x}_g)^{\mathrm{T}}\,\Sigma^{-1}\,(A_{\mathbf{w}}\mathbf{x}_g) = s_g\ . \tag{8}$$

To do that, we first note that for any pair of contrast matrices A_1 and A_2 with N_I columns and of rank $N_I - 1$, with each row sum equal to zero,

$$(A_1\mathbf{x}_g)^{\mathrm{T}}\,(A_1\Sigma A_1^{\mathrm{T}})^{-}\,(A_1\mathbf{x}_g) = (A_2\mathbf{x}_g)^{\mathrm{T}}\,(A_2\Sigma A_2^{\mathrm{T}})^{-}\,(A_2\mathbf{x}_g)\,.$$

Here a generalised inverse is used, defined as $BB^{-}B = B$, which gives

$$B^{-1} = B^{-}$$

when B is invertible. Now,

$$s_g = (A\mathbf{x}_g)^{\mathrm{T}}\,(A\Sigma A^{\mathrm{T}})^{-1}\,(A\mathbf{x}_g) = (A_{\mathbf{w}}\mathbf{x}_g)^{\mathrm{T}}\,(A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}})^{-}\,(A_{\mathbf{w}}\mathbf{x}_g)\ ,$$

so we can prove (8) by showing that

$$(A_{\mathbf{w}}\mathbf{x}_g)^{\mathrm{T}}\,\Sigma^{-1}\,(A_{\mathbf{w}}\mathbf{x}_g) = (A_{\mathbf{w}}\mathbf{x}_g)^{\mathrm{T}}\,(A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}})^{-}\,(A_{\mathbf{w}}\mathbf{x}_g)\,.$$

Since $A_{\mathbf{w}}$ is idempotent, this is the same as proving that

$$(A_{\mathbf{w}}\Sigma A_{\mathbf{w}})^{-} = A_{\mathbf{w}}^{\mathrm{T}}\Sigma^{-1}A_{\mathbf{w}}.$$

Writing $A_{\mathbf{w}}$ as

$$A_{\mathbf{w}} = I - \mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}$$

it follows that

$$\begin{aligned}
A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}}\left(A_{\mathbf{w}}^{\mathrm{T}}\Sigma^{-1}A_{\mathbf{w}}\right)A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}} =& A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}}\Sigma^{-1}A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}} \\
=& \left[I - \mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right]\Sigma\left[I - \mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right]^{\mathrm{T}}\Sigma^{-1} \\
&\times\left[I - \mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right]\Sigma\left[I - \mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right]^{\mathrm{T}} \\
=& \left[\Sigma - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right]\Sigma^{-1}\left[\Sigma - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right] \\
=& \Sigma - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}} = A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}}.
\end{aligned}$$

Thus,

$$(A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}})^{-} = A_{\mathbf{w}}^{\mathrm{T}}\Sigma^{-1}A_{\mathbf{w}}.$$

and (8) is proved.

Using this result, we can write the LRT as

$$\frac{|\bar{x}_g^w|}{\sqrt{s_g+2}} \geq k', \tag{9}$$

where $0 \leq k' < \infty$ is a new constant. To derive the distribution of the statistic that corresponds to (9) under the null hypothesis, we proceed as follows. Let

$$T_g = \sqrt{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}\,(N_I - 1 + 2\alpha)}\;\frac{\bar{X}_g^w}{\sqrt{S_g+2}}.$$

Then since

$$\bar{X}_g^w \sim \mathbf{N}\left(0, \frac{c_g}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right)$$

it can be shown that $\bar{X}_g^w$ is independent to all elements of $A_{\mathbf{w}}\mathbf{X}_g$ and thus to S_g. Furthermore,

$$T_g = \frac{\bar{X}_g^w/\sqrt{c_g/\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}}{\sqrt{S_g/c_g + 2/c_g}/\sqrt{N_I - 1 + 2\alpha}},$$

where the numerator is independent of S_g and has the same normal distribution conditionally on all c_g (and thus also unconditionally), showing that the denominator in this ratio expression is independent of the numerator. A similar argument shows that S_g/c_g and $2/c_g$ are independent, and since they are chi-square distributed with $N_I - 1$ and 2α degrees of freedom respectively, the sum is chi-square distributed with $N_I - 1 + 2\alpha$ degrees of freedom. Hence, under the null hypothesis, T_g is a t-distribution with $N_I - 1 + 2\alpha$ degrees of freedom,

$$T_g \mid \Sigma, \alpha \;\sim\; t_{N_I - 1 + 2\alpha}\;.$$

We call T_g the weighted moderated t-statistic.

References

D.J. Bakewell and E. Wit. Weighted analysis of microarray gene expression using maximum-likelihood. *Bioinformatics*, 21(6):723–729, 2005.

P. Baldi and A.D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.

M. Benson, L. Carlsson, M. Adner, M. Jernås, M. Rudemo, A. Sjögren, P.A. Svensson, R. Uddman, and Cardell L.O. Gene profiling reveals increased expression of uteroglobin and other anti-inflammatory genes in glucocorticoid-treated nasal polyps. *Journal of Allergy and Clinical Immunology*, 113(6):1137–1143, 2004.

D.-T. Chen. A graphical approach for quality control of oligonucleotide array data. *Journal of Biopharmaceutical Statistics*, 14(3):591–606, 2004.

S. Dudoit and J.Y.H. Yang. Bioconductor R packages for exploratory data analysis and normalization of cDNA microarray data. In G. Parmigiani, E.S. Garett, R.A. Irizarry, and S.L. Zeger, editors, *The Analysis of Gene Expression Data*. Springer, 2003.

C.I. Dumur, S. Nasim, A.M. Best, K.J. Archer, A.C. Ladd, V.R. Mas, D.S. Wilkinson, C.T. Garret, and A. Ferreira-Gonzalez. Evaluation of quality-control criteria for microarray gene expression analysis. *Clinical Chemistry*, 50(11):1994–2002, 2004.

R. Edgar, M. Domrachev, and A.E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

J.L. Hall, S. Grindle, X. Han, D. Fermin, S. Park, Y. Chen, R.J. Bache, A. Mariash, Z. Guan, S. Ormaza, J. Thompson, J. Graziano, S.E. de Sam Lazaro, S. Pan, R.D. Simari, and L.W. Miller. Genomic profiling of the human heart before and after mechanical support with a ventricular assist device reveals alterations in vascular signaling networks. *Journal of Physiological Genomics*, 17(3):283–291, 2004. URL `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GDS558`.

S. Hautaniemi, H. Edgren, P. Vesanen, M. Wolf, A.-K. Jarvinen, O. Yli-Harja, J. Astola, K. Olli, and O. Monni. A novel strategy for microarray quality control using bayesian networks. *Bioinformatics*, 19(16):2031–2038, 2003.

W. Huber, A. von Heydebreck, and M. Vingron. Analysis of microarray gene expression data. In M. et al. Bishop, editor, *Handbook of Statistical Genetics, 2nd Edition.* John Wiley & Sons, 2003.

R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15, 2003.

K. Johnson and S. Lin. QA/QC as a pressing need for microarray analysis: meeting report from CAMDA'02. *BioTechniques*, 34(suppl):S62–S63, 3 2003.

N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions Volume 2*. Wiley, 1995.

I. Lönnstedt and T. Speed. Replicated microarray data. *Statistica Sinica*, 12 (1):31–46, 2002.

J.S. Maritz. *Empirical Bayes Methods*. Methuen & Co. Ltd., 1970.

T. Park, S.-G. Yi, S. Lee, and J.K. Lee. Diagnostic plots for detecting outlying slides in a cDNA microarray experiment. *BioTechniques*, 38(3): 463–471, 2005.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`.

H. Robbins. An empirical Bayes approach to statistics. In J. Neyman, editor, *Third Berkeley Symposium on Mathematics and Probability*, pages 157–163, 1956.

C.P. Robert. *The Bayesian Choice*. Springer, 2003.

M.M. Ryan, S.J. Huffaker, M.J. Webster, M. Wayland, T. Freeman, and S. Bahn. Application and optimization of microarray technologies for human postmortem brain studies. *Nucleic Acids Research*, 55(4):329–336, 2004.

L. Shi, W. Tong, F. Goodsaid, F.W. Frueh, H. Fang, T. Han, J.C. Fuscoe, and D.A. Casciano. QA/QC: Challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Review of Molecular Diagnostics*, 4(6):761–777, 2004.

G.K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

G.K. Smyth, N.P. Thorne, and J. Wettenhall. *LIMMA: Linear Models for Microarray Data User's Guide*, 2003. URL `http://www.bioconductor.org`.

H. Tomita, M.P. Vawter, D.M. Walsh, S.J. Evans, P.V. Choudary, J. Li, K.M. Overman, M.E. Atz, R.M. Myers, E.G. Jones, S.J. Watson, H. Akil, and W.E. Bunney Jr. Effect of agonal and postmortem factors on gene expression profile: Quality control in microarray analyses of postmortem human brain. *Biological Psychiatry*, 55(4):346–352, 2004.

W. Tong, S. Harris, X. Cao, H. Fang, L. Shi, H. Sun, J. Fuscoe, A. Harris, H. Hong, Q. Xie, R. Perkins, and Casciano D. Development of public toxicogenomics software for microarray data management and analysis. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 549(1-2):241–253, 2004.

X. Wang, S. Ghosh, and S.W. Guo. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, 29(15): e75, 2001.

X. Wang, M.J. Hessner, Y. Wu, N. Pati, and S. Ghosh. Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics*, 19(11): 1341–1347, 2003.

Y.H. Yang, S. Dudoit, P. Luu, D. Lin, V. Peng, J. Ngai, and T.P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.