Multidimensional local false discovery rate for microarray studies

Alexander Ploner1,*, Stefano Calza2, Arief Gusnanto3 and Yudi Pawitan1

1Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden,
2Dipartimento di Scienze Biomiologie e Biotecnologie, Università degli Studi di Brescia, 11 25123 Brescia, Italy and
3MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR, UK

Received on August 29, 2005; revised on December 14, 2005; accepted on December 15, 2005

ABSTRACT

Motivation: The false discovery rate (fdr) is a key tool for statistical assessment of differential expression (DE) in microarray studies. Overall control of the fdr alone, however, is not sufficient to address the problem of genes with small variance, which generally suffer from a disproportionally high rate of false positives. It is desirable to have an fdr-controlling procedure that automatically accounts for gene variability.

Methods: We generalize the local fdr as a function of multiple statistics, combining a common test statistic for assessing DE with its standard error information. We use a non-parametric mixture model for DE and non-DE genes to describe the observed multi-dimensional statistics, and estimate the distribution for non-DE genes via the permutation method. We demonstrate this fdr2d approach for simulated and real microarray data.

Results: The fdr2d allows objective assessment of DE as a function of gene variability. We also show that the fdr2d performs better than commonly used modified test statistics.

Availability: An R-package OCplus containing functions for computing fdr2d() and other operating characteristics of microarray data is available at http://www.meb.ki.se/~yudpaw

Contact: alexander.ploner@meb.ki.se

1 INTRODUCTION

The extreme multiplicity of genes in microarray data has generated a keen awareness of the problem of false discoveries. Consequently, the concept of false discovery rate (fdr, Benjamini and Hochberg, 1995) has seen rapid development and wide-spread application to microarray data (e.g. Storey and Tibshirani, 2003; Reiner et al., 2003; Pawitan et al., 2005a). In addition to the multiple comparison issue, it is also well-known that conventionally significant genes with very small standard errors are more likely to be false discoveries (e.g. Tusher et al., 2001; Lönnstedt and Speed, 2002; Smyth, 2004). In current approaches, separate considerations—either ad hoc or model-based—are necessary to deal with genes with small standard errors, typically leading to a modified test statistic. In this paper, we treat both problems simultaneously by first generalizing the local fdr for multiple statistics, and then using this new concept to enhance the classical t-statistic in two-sample comparison problems.

We adopt a conventional mixture framework where a statistic Z is used to test the validity of the null hypothesis of no differential expression (DE) for each gene individually. Assuming that a proportion π0 of genes is truly non-DE, the density function f of the observed statistics is a mixture of the form

\[ f(z) = π_0 f_0(z) + (1 - π_0) f_1(z), \tag{1} \]

where \( f_d(z) \) is the density function of the statistics for non-DE genes and \( f_1(z) \) for DE genes. Although parametric mixture models have been used successfully in microarray data analysis (e.g. Newton et al., 2004), we will allow \( f() \), \( f_0() \) and \( f_1() \) to be non-parametric (c.f. Efron et al., 2001).

In this setting, the local false discovery rate—abbreviated as fdr to distinguish it from the global FDR as suggested by Benjamini and Hochberg (1995)—is defined as

\[ \text{fdr}(z) = π_0 \frac{f_d(z)}{f(z)}. \tag{2} \]

The local fdr can be interpreted as the expected proportion of false positives if genes with observed statistic \( Z = z \) are declared DE. Alternatively, it can be seen as the posterior probability of a gene being non-DE, given that \( Z = z \).

The key idea of this paper is that Equations (1) and (2) are immediately meaningful for a \( k \)-dimensional statistic \( Z = (Z_1, \ldots, Z_k) \). Specifically in the two-dimensional (2D) case \( k = 2 \), we can use two different statistics \( Z_1 \) and \( Z_2 \) that capture different aspects of the information contained in the data. We will refer to

\[ \text{fdr2d}(z_1, z_2) = π_0 \frac{f_d(z_1, z_2)}{f(z_1, z_2)}, \tag{3} \]

as the 2D local fdr and use it as a representative for higher dimensional fdrs. When necessary, the standard 1D fdr will be written explicitly as fdr1d.

We illustrate these concepts graphically for a simple simulated dataset of 10 000 genes and two groups of seven samples each, as described in detail in Section 2.4 (Scenario 1). This allows us to estimate all densities and fdrs involved directly from a second, much larger (107) set of simulations from the same underlying model, by keeping track of the known DE or non-DE status of each simulated gene, and counting the respective numbers in small intervals. This avoids the technical smoothing issues addressed in Section 2.1, and demonstrates the relevance of the fdr2d even in an idealized setting without gene–gene correlations or mean–variance relationships.

*To whom correspondence should be addressed.
For the common problem of comparing gene expression between two groups, a standard test statistic underlying the univariate local fdr is the classical t-statistic

\[ t_i = \frac{\bar{x}_{ij} - \bar{x}_{ij'}}{s_{ei}} \]

with pooled standard error

\[ s_{ei} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \sqrt{1/n_1 + 1/n_2}, \]

where \( \bar{x}_{ij} \), \( s_{ij} \), and \( n_j \) are the gene-wise group mean, standard deviation, and sample size for gene \( i \) and group \( j = 1, 2 \). We present the estimated densities \( f(z) \) and \( f_d(z) \) of the t-statistics for the simulated data in Figure 1a, and the corresponding univariate local fdr in Figure 1b.

Various extensions to 2D are possible, but for simplicity and to facilitate smoothing in real data situations, we include the standard error information directly in the fdr assessment by choosing

\[ z_{ij} = t_i \quad \text{and} \quad z_{ij'} = \log(||s_{ei}||), \]

where \( s_{ei} \) is the standard deviation of the group means.

The shape of the resulting densities \( f(z_1, z_2) \) and \( f_d(z_1, z_2) \) are indicated by the isolines in Figure 1c, and the resulting local fdr is shown by the isolines in Figure 1d. The sloped fdr isolines indicate that the 2D statistic \((t, \log \text{se})\) provides more information than the t-statistic on its own. The results confirm our expectation that genes with misleadingly small standard errors have higher fdr; e.g. for genes with log-standard error \( > -0.5 \), an absolute t-statistic of \( > 3.8 \) is sufficient for fdr2d \( \leq 0.05 \), whereas for genes with log-standard error \( < -1 \), an absolute t-statistic \( > 5.5 \) is required to achieve the same fdr. Thus, without extra modelling, the fdr2d offers an objective way to discount information from genes with misleadingly small standard errors.

To summarize our novel contributions, in this paper we introduce the concept of multi-dimensional local fdr, describe an estimation procedure for the 2D local fdr and demonstrate its application to simulated and real datasets; we show that without introducing any extra modeling or theoretical complexity, the fdr2d performs as well as or better than the previously suggested procedures in dealing with misleadingly small standard errors. Finally, we illustrate how the fdr2d can be transformed to make graphical representations like the volcano plot more objective.
2 MATERIALS AND METHODS

2.1 Estimating the local fdr

The non-parametric approach to fdr requires some smoothing operation for estimating the ratio of densities in Equation (2). We will start by presenting a feasible smoothing procedure in this section and go on to discuss the estimation of \( \pi_0 \) in the next section. Because of the practical difficulties of smoothing in higher dimensions, we will limit ourselves to the case \( k = 2 \). To be specific, we will consider the problem of two-independent-group comparison; more complex designs may require some modifications in the permutation step, but the smoothing problem is conceptually similar.

In the first step, we can generate the null distribution of \( Z \) by the permutation method, since under the null hypothesis of no difference, the group labels can be permuted without changing the distribution of the data (c.f. Efron et al., 2001). In order to allow general dependency between genes, each permutation is applied to all genes simultaneously. Let \( Z \) be the \( n \times k \) matrix of observed \( Z_s \) from \( n \) genes. Each permutation of group labels generates a new dataset and statistic matrix \( Z^* \). Let \( p \) be the number of permutations, so we have a series of \( Z^*_1, \ldots, Z^*_p \) representing samples of \( Z \) under the null. In all of the examples in this paper we use \( p = 100 \) permutations.

In principle, we could estimate \( f(z) \) using non-parametric density smoothing of the observed \( Z \), and similarly \( f_d(z) \) using \( Z^*_1, \ldots, Z^*_p \); then compute the \( fdr2d(z) \) by simple division. This approach, however, has two inherent problems: (1) because of the different amount of data, different smoothing is required for \( f(z) \) and \( f_d(z) \); and (2) optimal smoothing for the densities may not be optimal for the fdr. While these questions apply regardless of the dimension of the underlying statistic, we found them to be more consequential in the 2D than in the 1D setting.

We have therefore decided to implement a procedure that requires only a single smoothing of the ratio

\[
r(z) = \frac{p f_d(z)}{f(z) + p f_d(z)},
\]

and the estimated fdr is computed as

\[
fdr2d(z) = \pi_0 \frac{r(z)}{p(1 - r(z))}.
\]

(4)

The procedure is a 2D extension of the idea in Efron et al. (2001): in order to estimate \( r(z) \), we

1. Associate all the statistics generated under permutation with ‘successes’ and the observed statistics with ‘failures’. Then \( r(z) \) is the proportion of successes as a function of \( z \).
2. Perform a non-parametric smoothing of the success–failure probability as a function of \( z \).

Since the full dataset involving \( Z, Z^*_1, \ldots, Z^*_p \) is quite large, we pre-bin the data into small intervals and perform discrete smoothing of binomial data on the resulting grid. The smoothing is done using the mixed-model approach of Pawitan (2001, Section 18.10) as described below, and the resulting algorithm is fast despite the initial data size.

Let \( y_{ij} \) be the number of successes in the \((i, j)\) location of the grid and \( N_{ij} \) the corresponding number of points. By construction, \( y_{ij} \) is binomial with size \( N_{ij} \) and probability \( r_{ij} \), where the latter is a discretized version of \( r(z) \). We assume a link function \( h(\cdot) \) such that

\[
h(r_{ij}) = \eta_{ij},
\]

where \( \eta_{ij} \) satisfies the smoothness condition below. Given a smoothing parameter \( \lambda \), the smoothed estimate of \( r_{ij} \) is the minimizer of the penalized log-likelihood

\[
\log L(r, \lambda) = -\sum_{i,j} \left( y_{ij} \log r_{ij} + (N_{ij} - y_{ij}) \log (1 - r_{ij}) \right) + \lambda \sum_{(i,j) \in \mathcal{L}} (\eta_{ij} - \eta_{il})^2,
\]

where \((i, j) \sim (k, l)\) means that \((i, j)\) and \((k, l)\) are primary neighbors in the 2D lattice. The estimate is computed using the iteratively weighted least-squares algorithm (Pawitan 2001, Section 18.10), which is very stable in this case.

A practical complication arises from the presence of empty cells at the edges of the distribution. Naive treatment of these cells as having no event leads to serious bias. We reduce the bias by imputing single events to the empty cells, using the fact that genes with large \( t \)-statistics are likely to be regulated, and genes with \( t \)-statistics close to zero are likely to fall under the null hypothesis. We investigated in detail the logistic and identity link functions, and found the former to be more biased. Primarily, this is because the weights implied by the logistic-link function assign large influence to points in the center of the distribution; this creates a corresponding large bias in areas with low densities, which are precisely the region of interest for fdr assessment.

While it would be desirable to have an optimal and automatic smoothing parameter \( \lambda \), we have found that this is quite tricky to achieve. Although theory provides optimal smoothing for \( r(z) \), we face again the problem that what is optimal for \( r(z) \) is not necessarily optimal for \( fdr2d \). Instead, we can compute the effective number of parameters for the smooth (Pawitan, 2001, Section 18.10), which is limited by the number of bins in the grid. This allows the user to specify the desired degree of smoothness as a percentage, with lower values indicating stronger smoothing. The averaging property of \( fdr2d \) (Section 2.1.1) can be used to check whether the choice of smoothing parameter has been suitable, see Section 3.3.

To summarize, in practice we use the identity link function \( h(r_{ij}) = r_{ij} \), and the output of the smoothing step is a smooth estimate of \( r(z) \), evaluated at discrete points \((i, j)\). The \( fdr2d \) is then computed using (4) and a suitable estimate of \( \pi_0 \). In our experience, a relatively coarse grid on the order of \( 20 \times 20 \) will usually suffice; for these grids, we have found that relatively mild smoothing, allowing \( 70\%-80\% \) of the number of bins as the effective number of parameters of the smooth, is generally sufficient.

2.2 Estimating \( \pi_0 \)

The estimation of the proportion of non-DE genes is a pervasive problem when computing fdr’s (e.g. Storey and Tibshirani, 2003; Dalmasso et al., 2005; Pawitan et al., 2005b). We can obtain a conservative estimate by observing that, as a (conditional) probability, the \( fdr2d \) is bounded above by 1, so

\[
\pi_0 \leq \frac{f(z)}{f_0(z)}, \text{ for all } z
\]

so \( \pi_0 \leq \min_{z} f(z)/f_0(z) \), and we can take this upper bound as an estimate of \( \pi_0 \) (Efron et al., 2001). In practice, we do not recommend this approach for \( fdr2d \) even if we compute the bound only in the data-rich center of the distribution in order to account for possible instability in our estimate of \( f(z)/f_0(z) \) at the edges, we find that the bound is too dependent on the smoothing parameter.

Instead we suggest that \( \pi_0 \) be estimated from the \( fdr1d \). In principle, any procedure to estimate \( \pi_0 \) can be used. For example, this can be done via the upper-bound argument outlined above, which provides reasonably stable, though biased, estimates in the 1D case. A less biased estimate can be found using a mixture model for the \( t \)-statistics that we have described recently (Pawitan et al., 2005b). In this paper, we have used the upper-bound estimate for the simulated data—which is sufficiently accurate here—and the mixture-based estimates for the real datasets.
Fig. 2. fdr2d for a simulated dataset of 10,000 genes, based on two different estimates: (a) direct, using unsmoothed counts of the known DE and non-DE status of 10^7 genes simulated from the same model as the displayed data and (b) indirect, based only on the data displayed and without knowledge of the DE status of each gene, using permutations and smoothing as described in Section 2.1.

2.3 Monotonicity

Conditional on the log standard error, we make the estimated fdr2d decreasing with the absolute size of the observed test statistic by replacing it with a cumulative minimum:

\[
\text{fdr2d}(t, \log \text{se}) = \min_{0 \leq z \leq t} \text{fdr2d}(z, \log \text{se}) \quad \forall t > 0
\]

\[
\text{fdr2d}(t, \log \text{se}) = \min_{t \leq 0} \text{fdr2d}(t, \log \text{se}) \quad \forall t \leq 0.
\]

We have found that this requirement stabilizes the estimates at the margins of the observed distribution.

2.4 Simulation scenarios

We use five different simulation scenarios to demonstrate properties of the fdr2d. We assume 10,000 genes per array with a proportion of truly non-DE genes \(p_0 = 0.8\) throughout, and compare two independent groups with \(n = 7\) arrays per group. For Scenarios 1–4, we further assume that the log-expression values are also normally distributed in each group. In the simplest case of Scenario 1, both the mean difference for the DE genes and the gene-wise variances for all genes are fixed at \(D = \pm 1\) (with equal proportions) and \(\sigma^2 = 1\), respectively (Pawitan et al., 2005a). In contrast, both mean differences and gene-wise variances are random in Scenarios 2–4, which are based on the model described in Smyth (2004): for each gene \(i\), we simulate the variance from

\[
\sigma_i^2 \sim \frac{d_0 s_i^2}{X_i},
\]

and for each DE gene, we simulate the mean difference from

\[D_i \sim N(0, \nu_0 \sigma_i^2),\]

so that the absolute effect size \(|D_i|\) for each DE gene will be proportional to its variability. As in Smyth (2004), we set \(d_0 = 4\) and \(\nu_0 = 2\), and vary the degrees of freedom \(d_0\) to control the gene-wise variances:

- Scenario 2 has \(d_0 = 1000\), which results in very similar variances across genes,
- Scenario 3 has \(d_0 = 12\), which generates moderate variability in the variances,
- Scenario 4 has \(d_0 = 2\), which leads to strong variability of the variances and consequently very large \(t\)-statistics.

Note that Scenarios 2–4 correspond closely to the simulations with ‘similar’, ‘balanced’ and ‘different’ variances shown in Figures 2–4 in Smyth (2004).

In Scenario 5 finally we attempt to generate a realistic variance structure as present in real data. This is based on taking bootstrap samples from the BRCA dataset described in the following section: we take the original 3170 genes that were measured in two groups of \(n = 7\) and \(n = 8\) samples and subtract the corresponding group mean from the log-expression value for each gene. The resulting \(n = 15\) residuals are used as the plug-in estimate for the gene-wise error-distributions. We use the same bootstrap sample from the residuals for each gene, thereby preserving the correlations between the errors across genes. For genes randomly assigned to be DE, we simulate a log-fold change from \(N(1, 0.25)\) and add it to the observations in group 1. Hence the resulting dataset preserves to a large extent the distribution of variances and correlations in the underlying dataset.

2.5 Datasets

The BRCA data set was collected from patients suffering from hereditary breast cancer, who had mutations either of the BRCA1 gene \((n = 7)\) or the BRCA2 gene \((n = 8)\), as described in Hedenfalk et al. (2001). Expression was originally measured for 3226 genes, but following Storey and Tibshirani (2003), we removed 56 extremely volatile genes and analysed only the remaining 3170 genes. The mixture estimate \(\pi_0\) for this dataset is 0.61. (Note: the value \(\pi_0 = 0.43\) reported in Pawitan et al. (2005b) was based on the analysis of the raw scale, which is not as satisfactory as the log-scale used in the current paper).

We also applied our approach to the analysis of 240 cases of diffuse large B-cell lymphoma data described in Rosenwald et al. (2002), which was collected on a custom-made DNA microarray chip yielding measurements for 7399 probes. The average clinical followup for patients was 4.4 years, with 138 deaths occurring during this period. For this analysis, we ignored the censoring information and only compared 102 survivors with 138 nonsurvivors. The mixture estimate \(\pi_0\) for this dataset is 0.59 (Pawitan et al., 2005b).

2.6 Modified \(t\)-statistics

Previous attempts to overcome the small standard errors are based on an ad hoc fix by simply adding a constant to the observed standard error:

\[
t_i = \frac{x_{1i} - x_{2i}}{s_{c \hat{z}_i} + \hat{s}_0},
\]

where \(s_0\) is a fudge factor. Efron et al. (2001) simply computed \(s_0\) as the 90th percentile of the \(s_{c \hat{z}_i}\)‘s. In the so-called SAM method (Tusher et al., 2001), \(s_0\) is determined by a complex procedure based on minimizing the coefficient of variation of \(t_i\) as a function \(s_{c \hat{z}_i}\); for our computations, we used an
R implementation of SAM based on the siggenes package by Holger Schwender.

Alternatively, Smyth (2004) developed a hierarchical model for the gene-wise variance according to (5) and derived an empirical Bayes estimate of the gene-wise variance

\[
\hat{s}_i^2 = E(s_i^2 | s_i^2) = \frac{d_0 s_0^2 + d_1 s_i^2}{d_0 + d_1},
\]

where \(d_0\) and \(s_0^2\) are unknown parameters and \(d_1\) is the degrees of freedom associated with the sample variance \(s_i^2\). This variance estimate is used to get a moderated \(t\):

\[
\hat{t}_i = \frac{x_i - \bar{x}}{s_i \sqrt{1/n_1 + 1/n_2}},
\]

which under mild assumptions has a \(t\) distribution with \(d_0 + d_1\) degrees of freedom. Application of this method requires estimation of the hyperparameters \(d_0\) and \(s_0^2\) as described by Smyth and implemented in his R package limma.

### 3 RESULTS

#### 3.1 Simple properties of multidimensional fdr

The global FDR is the average of the local fdr, a useful relationship for characterizing a collection of genes declared DE by local methods. Suppose \(R\) is a rejection region such that all genes with (multi-dimensional) statistics \(z \in R\) are called DE. The global FDR associated with genes in \(R\) is

\[
FDR(R) = \frac{\pi_0 P(z \in R | H_0)}{P(z \in R)} = \int_R \frac{\pi_0 f_0(z)}{f(z)} \frac{f(z) P(z \in R)}{P(z \in R)} dz = E\{fdr(Z) | R\}.
\]

This means that the global FDR of gene lists found by fdr2d can be computed by simple averaging of the reported local fdr values, and consequently, fdr2d can be compared easily with other procedures in terms of its implied global FDR.

#### 3.1.1 Multidimensional fdr averages to 1D fdr

This first property is easy to demonstrate for a 2D vector \((z_1, z_2)\):

\[
fdr1d(z_i) = \pi_0 \frac{f_0(z_1)}{f(z_1)} = \pi_0 \frac{\int f_0(z_1, z_2)dz_2}{f(z_1)} = \pi_0 \frac{\int f_0(z_1, z_2) f(z_2 | z_1) dz_2}{f(z_1) f(z_2 | z_1)}
\]
\[
\int_{z_1} fdr2d(z_1, z_2) f(z_2 | z_1) dz_2 = E\{fdr2d(z_1, Z_2) | z_1\},
\]

i.e. the standard fdr1d(\(z_1\)) is an average of fdr2d(\(z_1, z_2\)) at a fixed value of \(z_1\). We found that oversmoothing of the fdr2d produces much smaller averages than expected from direct estimation of the fdr1d. This averaging property is therefore a useful calibration tool for checking that the 2D smoothing has been performed appropriately, see Figure 4.

### Multidimensional fdr is invariant under transformation

This property is useful if we consider different but mathematically equivalent choices of test statistics. Given a certain fdr2d(\(z\)), for any differentiable mapping \(u = g(z)\), the associated fdr is

\[
fdr2d_u(u) = \frac{fdr2d(g^{-1}(u))}{|g'(g^{-1}(u))|},
\]

i.e. no Jacobian is needed in the transformation. To show this, using obvious notations, if we account for the Jacobian of the transformation both for \(f_0\) and for \(f_z\):

\[
\begin{align*}
  f_{0u}(u) &= f_0(g^{-1}(u)) \left| \frac{\partial g^{-1}(u)}{\partial u} \right|, \\
  f_{zu}(u) &= f_z(g^{-1}(u)) \left| \frac{\partial g^{-1}(u)}{\partial u} \right|,
\end{align*}
\]

and the fdr2d\(_u\)(\(u\)) is the ratio of these densities, so the Jacobian terms cancel out.

While this is a pleasing quality in its own right, it has practical implications for the estimation and display of the fdr2d. Specifically, it allows us to use the most appropriate statistics for estimation and to transform the result into whatever scale desired. For example, to display the fdr2d as a function of the observed fold changes, we can choose

\[
z_{1i} = \bar{x}_{1i} - \bar{x}_{2i} \quad \text{and} \quad z_{2i} = \log_{10} \text{se}_i,
\]

i.e. the quantities underlying the \(t\)-statistic. Because of its characteristic shape for most microarray data, we might call this the ‘tornado plot’. An already established graphical display for studying the trade-off between effect size and significance is the volcano plot of \(\log_{10} P\)-values versus log-fold changes (Wolfinger et al., 2001), corresponding to

\[
z_{1i} = \bar{x}_{1i} = \bar{x}_{2} \quad \text{and} \quad z_{2i} = -\log_{10} P\text{-value}_i,
\]

Both the tornado and volcano plot contain inconvenient singularities—sharp edges—in the 2D distributions (Fig. 6), which make them hard to smooth. In contrast, the distribution of t-statistics and log-standard errors does not have this problem. We therefore estimate the fdr2d based on these statistics and use the invariance property to transform it as required.
3.2 Simulation study

Figure 2a shows the scatter plot of 10000 t-statistics and standard errors for a dataset simulated from Scenario 1 in Section 2.4, overlayed with the estimated fdr2d based on 10^7 genes simulated from the same model. For these extra simulations, we have kept track of the known DE status of each simulated gene, and count the proportion of false positives in each grid cell; no permutation or smoothing is involved, so the plot shows the true fdr2d function. In contrast, Figure 2b shows the fdr2d estimated from the dataset of 10000 genes using the method described in Section 2.1, without using extra simulations or extra knowledge of DE status. Comparison with Figure 2a shows that our proposed estimation method works as expected. Note also that even in this simple simulation setting, with no gene–gene correlations or mean–variance structures, the fdr2d isolines are still sloped.

We now compare the performance of fdr2d using the standard t-statistic versus fdr1d based on the standard and various moderated t-statistics described in Section 2.6: Efron's t (Efron et al., 2001), SAM (Tusher et al., 2001) and Smyth's t (Smyth, 2004). For each of the Scenarios 2–5 in Section 2.4, 100 datasets are generated, for a total of 10^4 simulated genes per scenario. For each fdr procedure, the genes are then ranked according to their true local fdr values, which are computed from the known DE status of each simulated gene, as above. In order to allow easy overall comparison, we compute the global FDR associated with the top ranking genes and draw the resulting FDR curve as a function of the proportion of genes declared DE.

The resulting FDR plots are given in Figure 3. The fdr2d with standard t shows top performance under all four scenarios. The fdr1d based on Smyth's moderated t performs equally well for Scenarios 2–4 (somewhat unsurprisingly, as these reflect its underlying model assumptions), but not as well for Scenario 5, where the error distribution is bootstrapped from the BRCA data. The fdr1d based on standard t does overall poorly and has actually worst global FDR in Scenarios 2 and 3; fdr1d based on Efron's t and SAM are mostly somewhere in between, though with dramatic breakdowns for specific scenarios: Efron's t does very badly in Scenario 4 under highly variable gene variances, and SAM has clearly inferior FDR in the BRCA-based Scenario 5. In summary, the fdr2d performs as well as the optimal empirical Bayes test for models with known variance structure, but better when the variance structure is close to the real BRCA data.

3.3 Application to real data

Results for the BRCA and Lymphoma data are shown in Figure 4 and Table 1. Figure 4a and b compare the fdr1d estimates (solid lines) with averaged fdr2d estimates (dashed lines). In both cases, there is overall good agreement between the curves, though the agreement is better in the tails of the curves than in the center, where the averaged fdr2d falls somewhat short of the fdr1d. This indicates that while the overall degree of smoothing is quite appropriate, there is still a problem with oversmoothing for t-statistics close to zero, fortunately an area of less practical interest.

Figure 4c and d show scatterplots of t-statistics versus log-standard errors, overlayed with estimated fdr2d isolines. For both datasets, we find pronounced asymmetry in these plots: not only in the amount of up- and down-regulation, which in any case is already evident from the fdr1d plots, but also in the way that genes with small standard errors are relatively discounted. In case of the BRCA data, the isoline for fdr2d = 0.05 is only very mildly sloped for positive t-statistics, leading to very moderate discounting of genes with log-standard errors < -2; the isoline for the same level in the left half of the plot is strongly curved, putting clearly more weight on genes with log-standard errors > -1. In case of the Lymphoma data, the asymmetry is even more pronounced, in that here genes with large standard errors are discounted for negative t-statistics, and genes with small errors for positive t-statistics (e.g. following the 0.1 isoline). This seems to reflect a strong mean–variance relationship for this dataset, which is also evident in the lower center of the plot.

Table 1 summarizes the number of genes that were found to be regulated for a fixed fdr cutoff (0.05 for the BRCA dataset, which has a stronger signal, and 0.1 for the Lymphoma set). For the BRCA data, exactly the same number of genes are found using fdr1d and fdr2d. For the Lymphoma data, however, the fdr2d finds both more genes to be regulated and also more balance between up- and down-regulation than fdr1d. This is a real example where the use of standard error information increases the power to detect DE. It is also interesting to note that the asymmetry in discounting genes based on their standard error can actually serve to find a better balance between up- and down-regulation.

More extensive comparisons between different procedures are summarized in Figure 5. In these comparisons, we also use the various modified t-statistics in the fdr2d procedure. In all cases, the fdr2d with the standard t-statistics performs as well as or better than any other procedure. As we go from 1D to 2D information, the advantage of the modified over the standard t-statistics disappears. This is an appealing practical property, as it means that we do not need to develop any extra analysis on how to modify the t-statistic.

3.4 Tornado and volcano plots

In Figure 6 we show tornado and volcano plots for the BRCA and Lymphoma data. The isolines are drawn the same as in Figure 4c and d, but were suitably transformed to the different scales. Both presentations show the same features that were discussed in the previous section, but as a function of the fold change instead of the t-statistic. Specifically, recall that the volcano plot is motivated by the same wish to discount significant genes with misleadingly small standard errors, i.e. genes with small fold change, yet significant P-value. The fdr2d allows us here to assess the plot object-
ively, and to define regions of interest that are more flexible than simple boxes and allow for asymmetry.

4 DISCUSSION

The generalization of the local fdr to multivariate statistics is conceptually simple, yet powerful enough to deal with two distinct problems in microarray data analysis: multiplicity and misleading over-precision. Both problems are well known, but we believe our approach is novel and potentially useful for most analysts. The technical problems in estimating the fdr, however, which are already present in the univariate case, become increasingly harder with the dimension of the test statistic. We have therefore limited our discussion to the 2D case, for which we have found the algorithm outlined in Section 2.1 to perform well.

Misleading over-precision is a subtle issue known in conditional inference (e.g. Lehmann, 1986, p. 558). Theoretically, this problem occurs even with a single test, although it is easier to explain with the confidence interval concept: short confidence intervals tend to have lower coverage than the stated confidence level (see Pawitan 2001, Section 5.10 for a simple example). Previously, awareness of this problem and biological intuition have lead to e.g. the volcano plot. The fdr2d provides a more flexible and objective control of the fdr for varying levels of information in the data.

We have applied the concept of fdr2d to the two-sample problem using a $t$-statistic. The classical $t$-statistic has been found to perform poorly on microarray expression data, especially for small datasets. Current solutions to this problem are either (1) ad hoc, such as the volcano plot or the test statistics proposed by Tusher et al. (2001) or Efron et al. (2001), or (2) make rather specific model assumptions, leading to empirical Bayes procedures such as the test statistics proposed by Lönnsted and Speed (2002) or Smyth (2004). In a specific dataset, the performance of the different procedures can vary considerably. The analytical solutions offer protection against very small standard errors inflating the $t$-statistic by modifying the denominator of the $t$-statistics. However, our simulation and data analysis suggest that ad hoc adjustments such as Efron’s $t$ can lead to poor performance if there is in fact significant variation between gene variances, which is expected in real data.

It is noteworthy that, without introducing any theoretical complexity, the fdr2d achieves comparable performance to the optimal empirical Bayes method. The fdr2d gains its power from acknowledging the mean–variance structure found in expression datasets of all sizes. The $t$-statistic reduces the combined information contained in the mean difference and the standard error to a single ratio. In contrast, the fdr2d uses the full bivariate information, with the effect that genes with $t$-statistics corresponding to very small standard errors are discounted for fdr reasons.

**Fig. 5.** Comparison between fdr2d based on standard $t$ and various other fdr procedures on the BRCA and lymphoma data. The curves represent global FDR as a function of the proportion of top-ranking genes declared DE by the different local fdr procedures. The solid line represents standard $t$ with fdr2d in all plots, the other lines stand for fdr1d in (a) and (b), and for fdr2d in (c) and (d). Dashes: standard $t$, dots: Efron’s $t$, dot-dashes: SAM, long dashes: Smyth’s moderated $t$.
Furthermore, we also found that there is no further gain in fdr2d from using modified t-statistics in its definition.

Various extensions of the fdr2d approach suggest themselves: for more than two groups of samples, we can split the classical F-statistic \( F = \frac{\text{MST}}{\text{MSE}} \) in the same manner as the t-statistic so that e.g.

\[
z_1 = \log \text{MST} \quad \text{and} \quad z_2 = \log \text{MSE}.
\]

Another possibility is the use of two different statistics that both test the same null hypothesis, but have different power against different alternatives, e.g. a t-statistic and a rank-sum statistic, comparable with the proposal made by Yang et al. (2005).

The two main problems that make the routine application of the fdr2d in data analysis difficult are the same as for the fdr1d, namely the smoothing of the ratio of densities and the estimation of \( p_0 \). These issues, however, are the subject of much ongoing research, which we believe will make possible a routine use of fdr2d.

ACKNOWLEDGEMENTS

The work of A.P. and Y.P. has been supported by a grant from the Swedish Cancer Society (Cancerfonden).

Conflict of Interest: none declared.

REFERENCES


