BMC Bioinformatics

A unified approach to false discovery rate estimation

Korbinian Strimmer (strimmer@uni-leipzig.de)

A unified approach to false discovery rate estimation

Korbinian Strimmer*[1]

[1] Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Härtelstr. 16–18, 04107 Leipzig, Germany

Email: Korbinian Strimmer*- strimmer@uni-leipzig.de;

*Corresponding author

Abstract

Background: False discovery rate (FDR) methods play an important role in analyzing high-dimensional data. There are two types of FDR, tail area-based FDR and local FDR, as well as numerous statistical algorithms for estimating or controlling FDR. These differ in terms of underlying test statistics and procedures employed for statistical learning.

Results: A unifying algorithm for simultaneous estimation of both local FDR and tail area-based FDR is presented that can be applied to a diverse range of test statistics, including p-values, correlations, z- and t-scores. This approach is semipararametric and is based on a modified Grenander density estimator. For test statistics other than p-values it allows for empirical null modeling, so that dependencies among tests can be taken into account. The inference of the underlying model employs truncated maximum-likelihood estimation, with the cut-off point chosen according to the false non-discovery rate.

Conclusions: The proposed procedure generalizes a number of more specialized algorithms and thus offers a common framework for FDR estimation consistent across test statistics and types of FDR. In comparative study the unified approach performs on par with the best competing yet more specialized alternatives. The algorithm is implemented in R in the "fdrtool" package, available under the GNU GPL from http://strimmerlab.org/software/fdrtool/ and from the R package archive CRAN.

Background

The false discovery rate (FDR) plays a prominent role in many high-dimensional testing and model selection procedures. Consequently, FDR methodologies are ubiquitous in the analysis of high-throughput data, such as in differential gene expression, SNP biomarker selection, peak detection in proteomic mass spectrometry data, or inference of edges in a network.

False discovery rate analysis starts with the seminal works by Schweder and Spjøtvoll [1] and by Benjamini and Hochberg [2]. Many others have followed suite, so that to date an impressive number of different algorithms for controlling and estimating false discovery rates have appeared in the literature.

In a nutshell, FDR estimation algorithms may be broadly categorized by the type of

- FDR,

- input test statistic, and

- employed inference procedures.

There are two main types of FDR, the "classic" tail area-based FDR (=Fdr) and local FDR (=fdr). Most FDR procedures are concerned either with Fdr or fdr, simultaneous estimation of both types of FDR is only possible with a few selected algorithms. With regard to test statistics, FDR calculations typically rely on p-values. However, FDR can be easily extended to other test statistics, such as correlations [3]. Relaxing the requirement of having p-values as input has the additional benefit that it allows for empirical null modeling [4]. Further key differences among the various FDR methods relate primarily to their respective procedures for density estimation and for inferring the proportion of null statistics.

Here, a unified statistical procedure for FDR estimation is described that generalizes a number of previous algorithms, specifically those of [5; 6; 4] and [7]. Notable features of thus approach include simultaneous estimation of Fdr and fdr from a diverse range of test statistics, its simplicity, very little a prior modeling assumptions, and the option of fitting the empirical null model.

The remainder of this paper is set out as follows. In the first part of the 'Methods' section a brief overview is given of the basic theory and definitions related to FDR and its estimation. In the second part of the 'Methods' section the proposed unified FDR procedure is described in detail. In the remaining part of the paper the new procedure is evaluated in comparison with other competing FDR estimation schemes.

Methods
Basic theory of FDR

This section gives a very brief review of the two component FDR model and the local and tail area-based FDR criteria. For a more refined discussion it is referred to [8] and references therein.

Throughout the paper the Efron naming conventions are followed. Specifically, "fdr" denotes the local false discovery rate, "Fdr" denotes the tail area-based false discovery rate, and "FDR" is a generic term encompassing both variants. Similarly, FNDR is the generic abbreviation for the false non-discovery rate [9].

In the following m simultaneous tests are considered, resulting in m test statistics such as $t_1, \ldots, t_m$ or $z_1, \ldots, z_m$ and corresponding p-values $p_1, \ldots, p_m$.

Tail area-based FDR

In order to control the number of false discoveries, i.e. the expected ratio $E(V/R)$ of the number of false positives V among all significant tests R, Benjamini and Hochberg [2] introduced the following linear step-up procedure. First, the p-values are ordered so that $p_{(1)} \leq \ldots \leq p_{(m)}$. Second, each value $p_{(i)}$ is compared with $q\frac{i}{m}$, where q is the desired FDR level. Finally, with $k = \max(i : p_{(i)} \leq q\frac{i}{m})$ all hypotheses belonging to $p_{(1)}, \ldots, p_{(k)}$ are rejected. [2] show that when the test statistics are independent then this procedure controls $E(V/R)$ at level $\leq q$.

The above procedure suggests the following simple correction of p-values, in the following called Benjamini-Hochberg (BH) rule:

$$p_i^{\mathrm{BH}} = p_i \frac{m}{\mathrm{order}(p_i)} \quad , i = 1, \ldots, m. \tag{1}$$

Here $\mathrm{order}(p_i)$ equals one for the smallest and m for the largest p-value, respectively. For comparison, the standard Bonferroni correction [10] is $p_i^{\mathrm{Bf}} = p_i m$, and hence $p_i \leq p_i^{\mathrm{BH}} \leq p_i^{\mathrm{Bf}}$.

A way to intuitively understand BH rule is to consider the following two-component mixture of the observed p-values,

$$
\begin{aligned}
f(p) &= \eta_0 f_0(p) + (1 - \eta_0) f_A(p) \\
&= \eta_0 + (1 - \eta_0) f_A(p).
\end{aligned}
\tag{2}
$$

For p-values the null density f_0 is the uniform distribution U(0,1) and corresponds to the "uninteresting" p-values, whereas f_A is an unspecified alternative density for the "interesting" p-values. This mixture

model may also be written in terms of distribution functions,

$$F(p) = \eta_0 F_0(p) + (1 - \eta_0) F_A(p)$$
$$= \eta_0 p + (1 - \eta_0) F_A(p). \tag{3}$$

Fig. 1 illustrates the p-value mixture model using the transformed statistic $y = 1 - p$.

This two-component model provides the means for *defining* the tail area-based false discovery rate "Fdr" and also the false non-discovery rate "Fndr"[9]. Specifically, $\text{Fdr}(p) = \eta_0 p / F(p)$ — see also Table 1. This "Bayesian" definition of "Fdr" [11] is closely related but not identical to the original approach by Benjamini and Hochberg. The key difference is that being density-based it implicitly assumes that the number of hypotheses is large ($m \to \infty$). Intriguingly, this allows to view most FDR procedures based on the observed test statistics as providing *estimates* of "Fdr" (note the subtle but important difference of estimating FDR versus controlling FDR).

In case of the BH-corrected p-values (Eq. 1), it turns out that this rule is simply the nonparametric empirical estimator of Fdr:

$$\text{Fdr}(p_i) := \text{Prob}(\text{"not interesting"} | P \le p_i)$$
$$= \frac{\eta_0 F_0(p_i)}{F(p_i)} = \frac{\eta_0 p_i}{F(p_i)}.$$

Plugging in the empirical cumulative density function (ECDF) $\hat{F}(p_i) = \frac{\text{order}(p_i)}{m}$ as estimator of $F(p)$ and using the conservative guess $\hat{\eta}_0 = 1$ yields

$$\widehat{\text{Fdr}}(p_i) = \frac{\hat{\eta}_0 p_i}{\hat{F}(p_i)} = p_i \frac{\hat{\eta}_0 m}{\text{order}(p_i)}$$
$$\le p_i \frac{m}{\text{order}(p_i)}.$$

It is instructive to compare the definitions of "Fdr" and "Fndr" for a given threshold y with those of "sensitivity" and "specificity" – see Table 1. Note that the order of conditioning is reversed in the two instances, but otherwise the definitions are very similar. Furthermore, both "Fdr-Fndr" and "sensitivity-specificity" offer the means for determining an optimal decision rule. In a conventional test situation the threshold y is chosen to maximize both sensitivity and specificity (i.e. typically specificity is fixed and power is maximized). Analogously, in an FDR analysis one seeks to minimize Fdr and Fndr (e.g, by fixing Fndr and minimizing Fdr). Hence, there is a tradeoff between Fndr and Fdr, just as there is a tradeoff between sensitivity and specificity. Note that the formal similarities between Fdr / Fndr and sensitivity / specificity is yet another reason for prefering the Bayesian variant of FDR over other more operational definitions.

4

The BH rule is popular due to its simplicity. However, often it is a rather conservative estimator of Fdr. One way to improve the BH rule is to substitute a more appropriate estimate of the null proportion η_0. This leads directly to the well-known q-values, which are refined BH estimates with various suggested options for the estimation of η_0 [12; 7].

Additionally, monotonicity is another issue where the BH rule is open to improvement. Specifically, direct application of the BH correction easily yields corrected p-values with a different ordering than that of the original test statistics. This unpleasant property has already been noted by [2], and indeed the "max" function in the original step-up procedure provides a corresponding fix (albeit a rather adhoc one). [5; 6] point out that this issue can be more elegantly resolved by requiring the distribution function $F(p)$ of the p-values to be concave and, correspondingly, the marginal density $f(p)$ to be monotonically decreasing. There many different ways for fitting the two component FDR mixture model (Eqs. 2 and 3) and for estimating densities and η_0. This explains the multitude of FDR approaches in existence. Common to all is some form of "zero assumption" to render the mixture model identifiable. Typically, for large p-values one assumes that there is no contamination with the alternative distribution, i.e. $\mathrm{Fndr}(p \to 1) = 0$ and therefore $f(p \to 1) = \eta_0$.

Local FDR

An alternative to the classic tail-area based FDR is the local FDR, abbreviated here as "fdr". Specifically, the local FDR is the probability of the null model conditioned the observed test statistic (see Table 1). Note that the local FDR takes is computed at the density level, in contrast to the Fdr that is based on cumulative densities.

This approach has mainly been advocated by Efron and a few others [13; 14; 15]. The key virtue of local FDR is that it is more readily interpretable than Fdr, as it is an empirical Bayesian posterior probability and not some variant of a corrected p-value. However, due to the use of densities it is also more difficult to estimate, in particular if the alternative distribution in the two-component model is not parametrically specified.

An important relation between Fdr and fdr is the property $\mathrm{Fdr}(p) \le \mathrm{fdr}(p)$ that holds if $\mathrm{fdr}(p)$ is monotonically decreasing with decreasing p-value.

Test statistics other than p-values and empirical null modeling

Virtually all FDR procedures – both local and tail area-based methods – are designed to work with p-values as input test statistics. Regardless the popularity of p-value-based approaches, in many instances it is often more beneficial to base the FDR calculations on the actual test statistic, such as on a regularized t-score, a z-score, or a correlation, rather than on a p-value.

The reason for this is as follows. Very often the theoretical null model is misspecified, due to dependencies among test statistics and other factors [16]. In turn, this may lead to overly pessimistic or too optimistic p-values, and thus to a violation of the implicit assumption of the FDR two-component model for p-values (namely that the null p-values are drawn from the uniform distribution). In such a case the resulting FDR values will also be biased, and thus unreliable.

Efron has shown that this can be elegantly avoided by retaining free parameters in the null model for the original test statistics (typically for location or scale) and estimating these parameters from the data [4]. Intriguingly, this *empirical null modeling* is greatly facilitated by high dimensions – and hence it is one of the few instances where high-dimensionality is not a curse but a blessing. There are various attempts to take account of the dependencies among p-values in FDR calculations, however it seems much more natural (and easier) to simply conduct Fdr and fdr calculations on the level of the original test statistic whilst employing an empirical null. In a recent paper these considerations are confirmed from a decision theoretic point of view [17].

Despite these apparent advantages empirical null modeling is currently available in only two FDR estimation algorithms, "locfdr" [4] and "fdrtool" (this paper). Note that fitting an empirical null is not tied to z-scores and the assumption of a normal null distribution, it is equally well feasible for any other test statistic, e.g., correlations [3].

Unified procedure for FDR estimation
Overview and motivation

From the discussion in the previous section it is clear that there exists a veritable range of FDR-related methods. An brief overview is given in Table 2, which lists thirteen FDR procedures for which an implementation for the R platform [18] is available.

The aim of this paper is to introduce an FDR estimation procedure that brings together many aspects that otherwise are only considered separately into one common and coherent setting. Thus, in a sense this offers a unified algorithm for FDR analysis.

Specifically, a procedure is proposed

- for the simultaneous estimation of both Fdr and fdr, regardless of the type of test statistic,

- that does not treat p-values any different from other test statistics,

- that maintains the ordering of original test statistics,

- that uses efficient and well tested techniques for estimating η_0 and null distribution,

- and that remains (largely) compatible with the well established "locfdr" and "qvalue" algorithms.

Furthermore, the algorithm is conceptually simple. Components in this scheme for Fdr/fdr analysis are a generalized definition of the test statistic, a non-parametric density estimator, an approach of fitting the null model, combined together in an effective fashion.

The present approach, discussed in detail in the following subsections, is best described as a marriage of the non-parametric Grenander approach of [5] and [6] with the empirical null modeling of [4]. An implementation is available in the R package "fdrtool" [19].

Generalized test statistic

Central to the algorithm is a generic definition of the underlying test statistic. Specifically, a statistic $y \geq 0$ is considered with properties such that large values of y indicate an "interesting" case, and, conversely, small values close to zero an "uninteresting" case. Examples for suitable statistics y include:

- $y = 1 - p$ where p is a p-value,

- $y = |z|$ where z is a normal score,

- $y = |r|$ where r is a correlation, and

- $y = |t|$ where t is a t-score.

Note that the choice of test statistic y automatically implies a corresponding null model $f_0(y; \theta)$, e.g., the uniform, half-normal, etc., which possibly contains some parameters θ. In terms of y the two-component model becomes

$$f(y) = \eta_0 f_0(y; \theta) + (1 - \eta_0) f_A(y) \tag{4}$$

7

and

$$F(y) = \eta_0 F_0(y; \theta) + (1 - \eta_0) F_A(y). \tag{5}$$

Accordingly, for a test statistic y the local FDR and the tail area-based FDR are given by

$$\mathrm{fdr}(y) = \eta_0 \frac{f_0(y; \theta)}{f(y)} \tag{6}$$

and

$$\mathrm{Fdr}(y) = \eta_0 \frac{1 - F_0(y; \theta)}{1 - F(y)}. \tag{7}$$

Furthermore, the p-value corresponding to the test statistic y equals $1 - F_0(y; \theta)$.

Density estimation using a modified Grenander approach

A central part of FDR analysis consists of the estimation of the marginal density $f(p)$ and the associated distribution function $F(p)$ from p-values p_i corresponding to the observed test statistics y_i.

The most simple approach is to use the empirical cumulative density function (ECDF) as estimator of $F(p)$. Note that the ECDF is the non-parametric maximum-likelihood estimate (NPMLE). The ECDF is very widely used in FDR analysis, including the two most popular FDR approaches (BH rule, "qvalue" algorithm). However, the ECDF has the disadvantage that it requires careful post-processing in order to achive monotone FDR values. Furthermore, it is a non-trivial issue to derive a density from the ECDF (see, e.g., [15] for an approach using loess smoothing). This important if computation of local FDR values is desired.

Another popular option, pursued in the "locfdr" program, is to estimate the density by spline Poisson regression on the histogram counts [21]. This work extremely well in general but can be problematic if the distribution has a strong peak – which is not uncommon, e.g., for p-values or partial correlations. Furthermore, this approach introduces additional parameters such as the histogram bin width or the degrees of freedom of the spline, which for some data may need diligent adjustment. In addition, as the approach does not place any monotonicity constraints on the density there is no guarantee that the order of the scores is maintained in the corresponding FDR values.

Other possibilities recently proposed for FDR density estimation include, e.g., normal mixtures [22], kernel-based approaches [23] and Bernstein polynomials [24].

An further alternative approach is provided by the Grenander density estimator [25]. In contrast to most other density estimators it has two main benefits which are highly useful in the context of FDR estimation: it explicitly incorporates monotonicity constraints (to preserve ordering of original test statistics) and provides simultaneous estimates of both PDF and CDF (to allow computation of both fdr and Fdr). Nonetheless, it is only slightly more complicated than than the ECDF. For FDR analysis the Grenander estimator has been first suggested by [5] and by [6].

Fig. 2 illustrates the mechanics behind Grenander density estimate. In essence, the Grenander density estimator is the decreasing piecewise-constant function equal to the slopes of the least concave majorant (LCM) of the ECDF. In the example shown in Fig. 2 the data x are $n = 30$ random samples from the exponential distribution with mean one. The left part of the figure shows the estimated monotonically decreasing density and the right part the corresponding empirical cumulative distribution. Note that the resulting distribution F is piecewise linear, whereas the density f is piecewise constant. The Grenander estimator is easy to obtain, as the LCM of the ECDF can be computed by monotone regression with weights [26]. Specifically, let x_i and y_i denote the coordinates for the ECDF, and $\Delta x_i = x_{i+1} - x_i$ and $\Delta y_i = y_{i+1} - y_i$. The slopes of the LCM are then given by antitonic regression of the raw slopes $\Delta x_i / \Delta y_i$ with weight Δx_i (see also the corresponding functions `monoreg`, `lcmgcm` and `grenander` in the "fdrtool" package). Like the ECDF, the Grenander estimator is also the NPMLE, with the added constraint of an underlying decreasing density.

Unfortunately, the standard Grenander estimator exhibits a severe shortcoming when applied the two-component FDR model: it leads to inconsistencies with the estimated η_0. This problem is extensively discussed in [5], and in fact causes these authors to abandon the Grenander estimator despite its favorable properties.

The point that is made here is that this deficiency can be easily fixed. Specifically, it is argued that the Grenander estimator is indeed very well suited for FDR analysis, but needs further modification in order to satisfy the additional constraints imposed by the two-component model.

The key problem can be understood best by going back to Eq. 3 which describes the FDR p-value mixture model on the level of the CDF. Effectively, this equation implies two constraints that any distribution compatible with the two-component model must satisfy:

- First, the CDF has to fulfill the condition $F(p) \geq \eta_0 p$ because $F(p) = \eta_0 p + (1 - \eta_0) F_A(p)$.

- Second, the condition $1 - F(p) \geq \eta_0(1 - p)$ must be met, because of

$$1 - F(p) = \eta_0(1 - p) + (1 - \eta_0)(1 - F_A(p)).$$

The second constraint is easy to overlook but is particularly important as Fig. 3 illustrates. There, it can be seen that the two-component model enforces a corridor of allowed values of the ECDF, where the width of this corridor depends on the parameter η_0. Note that the upper boundary (second constraint) ensures that the *minimum* possible slope equals η_0.

For FDR calculation this has the following consequence. The ECDF need not only be modified for monotonicity (Grenander estimator) but also need to be tailored to fit the constraints of the two-component model. This can be done as follows:

1. Compute the ECDF on the basis of the p-values.

2. Given η_0 impose the mixture model conditions on the ECDF for the p-values. Specifically, set $\hat{F}'(p_i) = \eta_0 p_i$ if $\hat{F}(p_i) < \eta_0 p_i$ (i.e. obey the lower boundary shown in Fig. 3) and likewise set $\hat{F}'(p_i) = 1 - \eta_0(1 - p_i)$ if $\hat{F}(p_i) > 1 - \eta_0(1 - p_i)$ (upper boundary).

3. The "modified" Grenander estimator is obtained as the standard Grenander estimator computed from the modified ECDF.

Note that the modified Grenander estimator retains the key property of the standard Grenander estimator (i.e. monotonicity) but in addition satisfies the constraints imposed by the two-component mixture model model. In particular, there are no inconsistencies with respect to the parameter η_0. This is illustrated in Fig. 4 where the modified Grenander estimator is applied with three different settings of η_0 to an example p-value data set. Note that by construction the modified Grenander density equals η_0 for large p-values.

Estimation of null sub-density by truncated maximum-likelihood

For computing p-values and the modified Grenander density suitable estimates of the parameters θ and η_0 are required. In other words, the null sub-density $\eta_0 f_0(y; \theta)$ of the two-component model (Eqs. 4 and 5) needs to be fit to the observed test statistics. This is straightforward in fully parametric models such as BUM [27]. However, it is often preferred to leave $f_A(y)$ unspecified. As a consequence, standard procedures for inferring mixture models such as the EM algorithm cannot be applied.

Instead, a truncated maximum-likelihood approach is applied here. In more detail, in this method the data are *censored* at some threshold y_c, so that only test statistics $\mathbf{y}_t = \{y_i : y_i < y_c\}$ are retained. The underlying assumptions is that for $y_i < y_c$ (nearly) all data points belong to the null part. This is called

the "strong zero assumption" in [28]. The truncated null density becomes $f_0^t(y;\theta) = f_0(y;\theta)/F(y_c;\theta)$ for $y < y_c$ and $f_0^t = 0$ otherwise. Maximization of the corresponding likelihood function returns $\hat{\eta}_0$ as well as an estimate of its asymptotic error. Similarly, the proportion of null values η_0 is inferred by assuming a binomial model for the observed number m_t of hypotheses in the set $\mathbf{y}_t$, which leads to the simple estimate $\hat{\eta}_0 = \max\{1, \frac{m_t}{m}/F(y_c;\hat{\theta})\}$ plus an associated error.

Truncated maximum-likelihood is the basis of the "locfdr" MLE algorithm [29; 28]. If the test statistics are p-values then the truncated maximum-likelihood algorithm reduces to the simple cutoff technique used in "qvalue" and most other p-value-based packages.

Selection of suitable truncation point using the false non-discovery rate

Fitting the null model and the associated parameters θ and η_0 by truncated maximum-likelihood depends on the choice of a suitable cutoff point y_c. In general, one wishes to select y_c such that the threshold is small enough to ensure that the zero assumption is met and that there is minimal bias due to contamination with the alternative $f_A(y)$. On the other hand, y_c needs be chosen large enough so that the number of data points in $\mathbf{y}_t$ is sufficient for reliably estimating θ and η_0.

The default "smoothing" approach employed in "qvalue" specifically aims at achieving an unbiased estimate of η_0 [7]. This is obtained by varying the cutoff point between zero and one, and subsequently estimating η_0 by interpolation at $y_c = 1$, i.e. for complete censoring.

For empirical null modeling the choice of an optimal y_c is more complicated. Table 3 lists the algorithms employed in various versions of the "locfdr" algorithm. Essentially, "locfdr" either uses a fixed y_c or it relies on a heuristic analytical formula aimed at reducing the mean squared error of the null sub-density [28]. Both approaches are not straightforward to extend to arbitrary test statistics y_i. Instead, here a more simple alternative procedure is proposed that enforces the "zero assumption" by requiring that the false non-discovery rate (Fndr) is minimized (i.e. y_c is chosen such that $\text{Fndr}(y_c)$ is small). Intriguingly, this leads to the following circular inferential problem: in order to determine a suitable cutoff y_c the Fndr must be known, yet to compute Fndr and other FDR quantities a suitable value for y_c must be specified. Fortunately, for most data sets the location of the cutoff point y_c need only be known approximately.

The "FNDR" strategy employed in "fdrtool" proceeds in two stages. In the first step, the mixture model is fit approximately, which leads to an approximate Fndr curve from which an approximately optimal y_c is obtained. In the second step truncated maximum-likelihood estimation on the basis of the approximate cutoff y_c is used for a refined fit of the mixture model which in turn allows to compute FDR quantities of

11

interest.

A simple approximate fit of the null model is achieved by matching its median $F_0^{-1}(1/2; \theta)$ with that of the observed y_i. Thus, a robust estimate of scale is used just as in the "locfdr" algorithm, see Table 3 (note that the median for the half-normal distribution corresponds to the interquartile range (IQR) of the corresponding normal with mean zero). Subsequently, after converting the test statistics into p-values an approximate estimate of the null proportion is determined by estimating η_0 for various cutoff-points and finally settling for the 0.1 quantile of the resulting distribution of corresponding η_0.

In addition to selecting y_c by the above "FNDR" approach, further methods available in the "fdrtool" package include the "locfdr" cutoff method [28] and the specification of the fraction of data points to be considered for estimating the empirical null. In a practical analysis it is always advisable to conduct the FDR calculations for various choices of y_c, (even though truncated maximum-likelihood appears to be fairly robust with regard to y_c).

Gluing it all together

With the above preliminaries, a general algorithm for estimating Fdr and fdr from an arbitrary test statistics y_i can be put together as follows:

1. Determine a suitable truncation point y_c.

2. Estimate the null model and its parameters, yielding $\hat{\eta}_0$ and $\hat{\theta}$.

3. Convert test statistics into p-values via $p_i = 1 - F_0(y|\hat{\theta})$.

4. Estimate the PDF $\hat{f}_p(p)$ and CDF $\hat{F}_p(p)$ of the p-values using the modified Grenander estimator (note that this requires $\hat{\eta}_0$).

5. Compute estimates of Fdr and fdr values based on p-values:

$$\widehat{\text{fdr}}_p(p) = \frac{\hat{\eta}_0}{\hat{f}_p(p)}$$

$$\widehat{\text{Fdr}}_p(p) = \frac{\hat{\eta}_0 p}{\hat{F}_p(p)}$$

6. Compute estimated Fdr and fdr values as a function of the original test statistics y:

$$\widehat{\text{fdr}}(y) = \widehat{\text{fdr}}_p(1 - \hat{F}_0(y))$$

$$\widehat{\text{Fdr}}(y) = \widehat{\text{Fdr}}_p(1 - \hat{F}_0(y))$$

7. Compute CDF and PDF on the y-scale:

$$\hat{f}(y) = \hat{\eta}_0 \frac{\hat{f}_0(y)}{\widehat{\text{fdr}}(y)}$$

$$\hat{F}(y) = 1 - \hat{\eta}_0 \frac{1 - \hat{F}_0(y)}{\widehat{\text{Fdr}}(y)}$$

Note that this transformation is directly derived from the definition of fdr and Fdr in Eqs. 6 and 7.

8. Estimate alternative sub-density:

$$\hat{F}_A(y) = \frac{\hat{F}(y) - \hat{\eta}_0 \hat{F}_0(y)}{1 - \hat{\eta}_0}$$

$$\hat{f}_A(y) = \frac{\hat{f}(y) - \hat{\eta}_0 \hat{f}_0(y)}{1 - \hat{\eta}_0}$$

Results and discussion
Computer simulations for p-value-based analyses

In order to assess the performance of the "fdrtool" algorithm it was compared with a number of other FDR procedures. Specifically, the R packages "fdrtool" version 1.2.4, "qvalue" version 1.1 [7], "locfdr" version 1.1-6 [4], "twilight" version 1.14.1 [30], "kerfdr" version 1.0.0 [23] and "nFDR" version 0.0 [24] were investigated.

First, FDR approaches based on p-values were studied. As generative model p-values were simulated from a mixture of the uniform $U(0, 1)$ with either the truncated exponential density $f_A(p; a) = \frac{a}{\exp(a)-1} \exp(a(1-p))$ or with the uniform $f_A(p; a) = U(0, a)$. Sample size and mixture model parameter a were varied, and from each generated data set the proportion of null p-values, and the squared error of the local FDR and the tail area-based FDR was estimated. The references for computing the squared error were the theoretical Fdr and fdr values derived from the assumed model.

Fig. 5 displays the results for three different cases: "model 1" is a uniform-exponential mixture with $\eta_0 = 0.8$ and $a = 5$, "model 2" is identical to "model 1" except for $a = 20$, and "model 3" utilizes the uniform-uniform mixture with $a = 0.2$. In all cases there were $B = 1000$ repeats and the number of p-values was $m = 200$.

The first column of Fig. 5 shows the accuracy in estimating η_0. Overall, the "kerfdr" and the "fdrtool" algorithms exhibit the smallest variability at the expense of a slightly biased estimate of η_0, especially for "model 1". In contrast, "qvalue" always produces nearly unbiased estimates but has a much higher variance. The "twilight" and "nFDR" are similar to "qvalue", but are less variable.

The second and third columns of Fig. 5 depict the error in the actually estimated Fdr and fdr values for various algorithms under the three model scenarios. In terms of correctly estimating fdr values all investigated packages with capability of computing local FDR (i.e. "fdrtool", "kerfdr", and "twilight") perform roughly equally well across all scenarios For "model 3" the "fdrtool" appears to have a slight advantage over the competing approaches. When comparing the accuracy of Fdr values "fdrtool" outperforms both "qvalue" and "nFDR", even though the differences are not large. The squared error of Fdr computed by "qvalue" and by "nFDR" exhibits more extreme outliers than those of "fdrtool".

Simulations and analysis of gene expression data for z-scores

In a further simulation study estimation of FDR from z-scores was considered with empirical null modeling. Specifically, data were simulated from a mixture of the normal distribution $N(\mu = 0, \sigma = 2)$ with the symmetric uniform alternatives $U(-10, -5)$ and $U(5, 10)$.

An example of the results from the simulations for sample size $m = 200$, $B = 1000$ repeats and $\eta_0 = 0.8$ is shown in Fig. 6. The estimates of the mixing parameter η_0 and of the scale parameter σ are slightly upwardly biased in "locfdr" but more importantly they also exhibit larger variability compared to "fdrtool". The mean squared error of the fdr vales are similar for both algorithms. Note that in this simulation a 75% quantile cutoff point was employed in "fdrtool" for the truncated maximum-likelihood estimation of the null model. The third row of Fig. 6 shows the FDR errors for z-scores with absolute values larger than 2. In this domain both investigated algorithms perform again very similar, but the average error is larger in comparison to the situation when all z-scores are included in the analysis.

In order to further compare the empirical null modeling a HIV and a breast cancer (BRCA) microarray gene expression data set was reanalyzed, following [31] and [4]. For the detailed biological background and the experimental setup it is referred to the original papers.

The HIV data consists of 7680 z-scores. The fit of "fdrtool" to the median-centered data is shown in Fig. 7. Specifically, the standard deviation of the null normal density was estimated to be $\hat{\sigma} = 0.786$ and the mixing parameter $\hat{\eta}_0 = 0.9575$. The number of discoveries with a fdr value smaller than 0.2 was 119. The "locfdr" algorithm finds a very similar null model, namely $\hat{\sigma} = 0.754$ and $\hat{\eta}_0 = 0.9342$. Corresponding to the smaller $\hat{\eta}_0$ "locfdr" detects 160 significant z-scores with fdr < 0.2.

The breast cancer data has size 3226. After median-centering the data were again supplied to both the "fdrtool" and "locfdr" packages. Both algorithms indicated that there was overdispersion ($\hat{\sigma} = 1.51$ versus $\hat{\sigma} = 1.55$) and and the proportion of null values was estimated to be $\hat{\eta}_0 = 1$. Correspondingly, for the

BRCA data there were no significant z-scores (note that this is in contrast to claims otherwise in [31]). In short, "fdrtool" and "locfdr" provide very similar analyzes both in terms of empirical null estimation and inferred fdr values.

Computational efficiency

Finally, the investigated FDR procedures were also compared in terms of computational efficiency. The (by a large margin) slowest program is "twilight". In contrast, the fastest algorithms are "fdrtool", "locfdr" and "qvalue", followed by "kerfdr" and "nFDR".

Conclusions

False discovery rate analysis is a key statistical innovation that has found widespread application in the study of high-dimensional data. One of the intriguing aspects of FDR is that can be understood both from a frequentist and Bayesian perspective. This has lead to a plethora of FDR criteria and FDR inference procedures.

The goal for the development of the "fdrtool" procedure was to establish a common framework that brings together the most compelling features of existing FDR methods. Specifically, novel features of the proposed "fdrtool" algorithm include

- a unified treatment of p-values and other test statistics, with identical algorithms and learning procedures,

- simultaneous and coherent estimation of both Fdr and fdr,

- empirical null modeling for test statistics other than z-scores,

- a method for selecting the truncation point based on controlling FNDR, and

- a simple semiparametric model using a modified Grenander density estimator.

Hence, "fdrtool" allows to compute local FDR values from p-values but also Fdr values from z-scores while taking account of an empirical null model. Despite the generality of the algorithm, it was shown that the accuracy of the algorithm is on par with the best competing yet more specialized FDR procedures. Moreover, the modular structure of the "fdrtool" procedure facilitates future extensions.

In summary, the "fdrtool" package and algorithm constitutes a comprehensive and feature-rich tool for a wide range of FDR-type analyzes.

15

During revision a referee pointed out that the distribution of observed p-values might be U-shaped [20]. This occurs, among other possibilities, if the null model is misspecified. As a result, the computed null p-values do not follow a uniform distribution, and thus by definition are improper. "fdrtool" cannot be applied directly to improper p-values, however, in these instances it might instead be preferable to conduct the FDR analysis on the level of the original test statistics.

Acknowledgements

References

1. Schweder T, Spjøtvoll E: **Plots of p-values to evaluate many tests simultaneously**. *Biometrika* 1982, **69**:493–502.

2. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *J. R. Statist. Soc. B* 1995, **57**:289–300.

3. Schäfer J, Strimmer K: **An empirical Bayes approach to inferring large-scale gene association networks**. *Bioinformatics* 2005, **21**:754–764.

4. Efron B: **Large-scale simultaneous hypothesis testing: the choice of a null hypothesis**. *J. Amer. Statist. Assoc.* 2004, **99**:96–104.

5. Langaas M, Lindqvist BH, Ferkingstad E: **Estimating the proportion of true null hypotheses, with application to DNA microarray data**. *J. R. Statist. Soc. B* 2005, **67**:565–572.

6. Broberg P: **A comparative review of estimates of the proportion unchanged genes and the false discovery rate**. *BMC Bioinformatics* 2005, **6**:199.

7. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc. Natl. Acad. Sci. USA* 2003, **100**:9440–9445.

8. Efron B: **Microarrays, empirical Bayes, and the two-groups model**. *Statistical Science* 2008, **23**:to appear.

9. Genovese C, Wassermann L: **Operating characteristics and extensions of the false discovery rate procedure**. *J. R. Statist. Soc. B* 2002, **64**:499–517.

10. Bonferroni CE: **Il calcolo delle assicurazioni su gruppi di teste**. In *Studi in Onore del Professore Salvatore Ortu Carboni*, Rome 1935:13–60.

11. Storey JD: **The positive false discovery rate: A Bayesian interpretation and the q-value**. *Ann. Statist.* 2003, **31**:2013–2035.

12. Storey JD: **A direct approach to false discovery rates**. *J. R. Statist. Soc. B* 2002, **64**:479–498.

13. Efron B, Tibshirani R, Storey JD, Tusher V: **Empirical Bayes analysis of a microarray experiment**. *J. Amer. Statist. Assoc.* 2001, **96**:1151–1160.

14. Efron B: **Robbins, empirical Bayes, and microarrays**. *Annals of Statistics* 2003, **31**:366–378.

15. Aubert J, Bar-Hen A, Daudin JJ, Robin S: **Determination of the differentially expressed genes in microarray experiments using local FDR**. *BMC Bioinformatics* 2004, **5**:125.

16. Efron B: **Correlation and large-scale simultaneous significance tesing**. *J. Amer. Statist. Assoc.* 2007, **102**:93–103.

17. Sun W, Cai TT: **Oracle and adaptive compound decision rules for false discovery control**. *J. Amer. Statist. Assoc.* 2007, **102**:901–912.

18. R Development Core Team: **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria 2007, [http://www.R-project.org]. [ISBN 3-900051-07-0].

19. Strimmer K: **fdrtool: a versatile R package for estimating local and tail area-based false discovery rates**. *Bionformatics* 2008, **24**:1461–1462.

20. Pounds S, Cheng C: **Robust estimation of the false discovery rate**. *Bioinformatics* 2006, **22**:1979–1987.

21. Efron B, Tibshirani R: **Using specially designed exponential families for density estimation**. *Ann. Statist.* 1998, **24**:2431–2461.

22. McLachlan GJ, Bean RW, Jones LBT: **A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays**. *Bioinformatics* 2006, **22**:1608–1615.

23. Robin S, Bar-Hen A, Daudin JJ, Pierre L: **A semi-parametric approach for mixture models: application to local false discovery rate estimation.** *Comput. Statist. Data Analysis.* 2007, **51**:5483–5493.

24. Guan Z, Wu B, Zhao H: **Nonparametric estimator of false discovery rate based on Bernstein polynomials**. *Statistica Sinica* 2008, **18**:in press.

25. Grenander U: **On the theory of mortality measurement, Part II**. *Skan. Aktuarietidskr.* 1956, **39**:125—153.

26. Robertson T, Wright FT, Dykstra RL: *Order restricted statistical inference*. John Wiley and Sons 1988.

27. Pounds S, Morris SW: **Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values**. *Bioinformatics* 2003, **19**:1236–1242.

28. Turnbull BB: **Optimal estimation of false discovery rates**. Tech. rep., Stanford University 2007, [http://www.stanford.edu/~bkatzen/optimal-FDR.pdf].

29. Efron B: **Size, power and false discovery rates**. *Ann. Statist.* 2007, **35**:1351–1377.

30. Scheid S, Spang R: **A stochastic downhill search algorithm for estimating the local false disovery rate**. *IEEE T. Comp. Biol. Bioinf.* 2004, **1**:98–108.

31. Jin J, Cai TT: **Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons**. *J. Amer. Statist. Assoc.* 2007, **102**:495–506.

32. Dalmasso C, Bröet P, Moreau T: **A simple procedure for estimating the false discovery rate**. *Bioinformatics* 2005, **21**:660–668.

33. Liao JG, Lin Y, Selvanayagam ZR, Shih WJ: **A mixture model for estimating the local false discovery rate in DNA microarray analysis**. *Bioinformatics* 2004, **20**:2694–2701.

Figure legends
Figure 1:

Two-component mixture model for p-values with cutoff point y_c. This implies a decision rule with errors α_1 and α_2. Further abbreviations: FP, false positives; TP, true positives; FN, false negatives; TN, true negatives. Note that here these quantities are all fractions (not counts) and that FP+TP+FN+TN $= 1$.

Figure 2:

Illustration of the Grenander density estimator. Left: Grenander density estimate (blue line); right: the corresponding concave distribution function (blue line) and the underlying ECDF (thine black lines).

Figure 3:

Constraints imposed by the p-value mixture model on the ECDF of p-values. The lower diagonal line corresponds to the constraint $F(p) \geq \eta_0 p$ whereas the upper diagonal line represents the constraint $1 - F(p) \geq \eta_0(1 - p)$. Right: the unmodified ECDF; left: the ECDF subject to the constraint $\eta_0 = 0.7$.

Figure 4:

The modified Grenander estimator computed from p-values for three different choices of η_0. Note that the underlying data are the same in all three instances and that the density of the modified Grenander equals η_0 for large p-values.

Figure 5:

Comparison of the estimates of the proportion of null p-values and the squared error of Fdr and fdr for various p-value-based FDR estimation procedures under three different simulation scenarios. The box plots summarize the estimates from $B = 1000$ repetitions of the simulations. The sample size (i.e. the number of multiple tests) was $m = 200$.

Figure 6:

Accuracy and variability of estimates of the null proportion η_0 and the scale parameter σ as well as the squared error of Fdr and fdr for z-score based algorithms. As in Fig. 5 $m = 200$ and $B = 1000$ was used. For the plots in the third row only z-scores with $|z| > 2$ were used.

Figure 7:

Graphical output provided by "fdrtool" for the HIV data set. The first row shows the densities, the second the distribution function and the last row the local and tail area-based false discovery rates.

Table 1: Definitions of FDR quantities contrasted with that of specificity and sensitivity.

Quantity	Definition	
Fdr	$= \mathrm{Prob}(\text{``not interesting''}|Y \geq y)$	$= \frac{FP}{TP+FP}$
Fndr	$= \mathrm{Prob}(\text{``interesting''}\;|Y \leq y)$	$= \frac{FN}{FN+TN}$
Sensitivity, Power	$= \mathrm{Prob}(Y \geq y|\text{``interesting''})$	$= \frac{TP}{TP+FN}$
Specificity	$= \mathrm{Prob}(Y \leq y|\text{``not interesting''})$	$= \frac{TN}{TN+FP}$
fdr	$= \mathrm{Prob}(\text{``not interesting''}|Y = y)$	
fndr	$= \mathrm{Prob}(\text{``interesting''}\;|Y = y)$	$= 1 - \mathrm{fdr}$

Note that in the above definitions TP, FP, FN, FP are assumed to be the true fractions - see Fig. 1. If instead observed counts are used, additionally the expectation $E(.)$ must be taken, conditioned on seeing a non-zero denominator (cf. [7; 11]).

Table 2: Overview over some commonly used FDR estimation procedures.

Name	FDR Type	Input Data	Comments	Description of Algorithm
fdrtool	fdr, Fdr	p-values, z-scores, t-scores, correlations.	Modified Grenander density , estimate, empirical null model, selection of truncation point by FNDR.	This paper
BUM	fdr, Fdr	p-values	Completely parametric model.	[27]
SAGx	fdr, Fdr	p-values	Grenander density estimate.	[5; 6]
qvalue	Fdr	p-values	Diverse estimates of η_0 available.	[12; 7]
nFDR	Fdr	p-values	Bernstein polynomial density.	[24]
multtest	Fdr	p-values	Benjamini-Hochberg algorithm.	[2]
LBE	Fdr	p-values	Location-based estimator.	[32]
locfdr	fdr	z-scores	Regression spline density estimate, empirical null model.	[4; 28]
nomi	fdr	z-values	Normal mixture modeling of non-null density.	[22]
LocalFDR	fdr	p-values	Uses loess smoothing.	[15]
kerfdr	fdr	p-values	Kernel density estimator.	[23]
twilight	fdr	p-values	KS fit of truncation point.	[30]
localFDR	fdr	p-values	Based on stochastic order model.	[33]

Note: The first column refers to an R package or R script implementing the respective FDR algorithm.

Table 3: Various choices of normal truncation points implemented in "locfdr".

Version	Released	Truncation Point	Reference
1.1-1	July 2006	$z_0 = 2$	
1.1-3	December 2006	$z_0 = \hat{\mu} + b\hat{\sigma}$ with $b = 3.55 - 0.44\log_{10}(m)$, $\hat{\mu} = \text{median}(z_i)$ and $\hat{\sigma} = \text{IQR}(z_i)/1.349$.	
1.1-6	November 2007	as version 1.1-3, but with $b = \max(1, 4.3m^{-0.112966})$	[28]

Note that $\hat{\mu}$ and $\hat{\sigma}$ are robust location and scale estimates, respectively.

"interesting"
(=alternative)

"not interesting"
(=null)

TP

FN α_2

TN α_1

FP

η_0

0   y   1

"not significant"   "significant"

$y = 1 - \text{pvalue}$
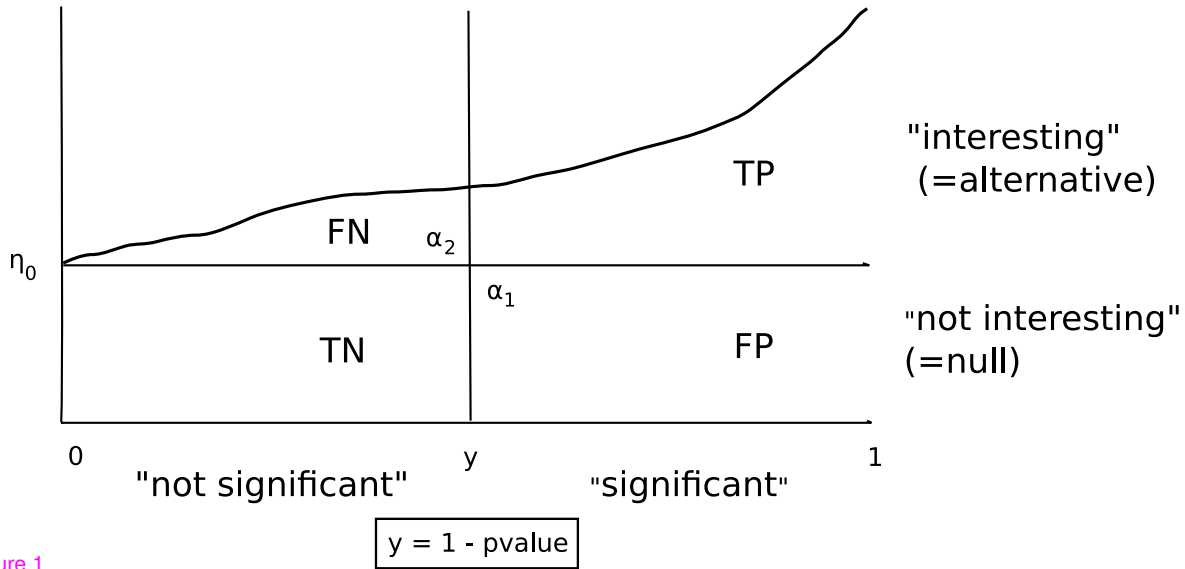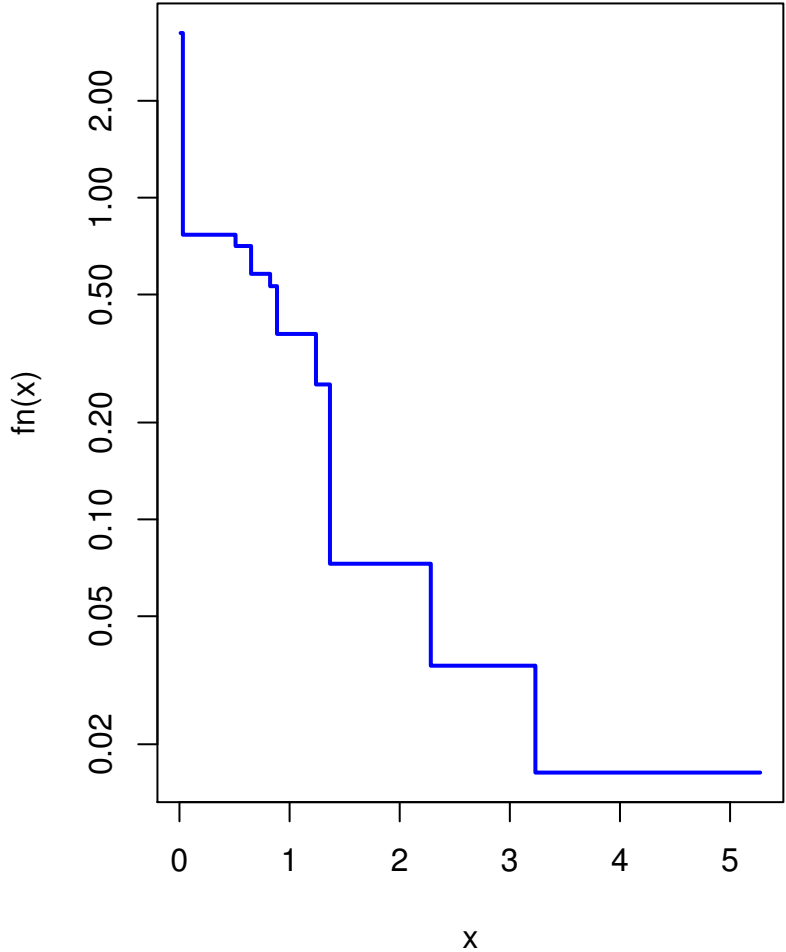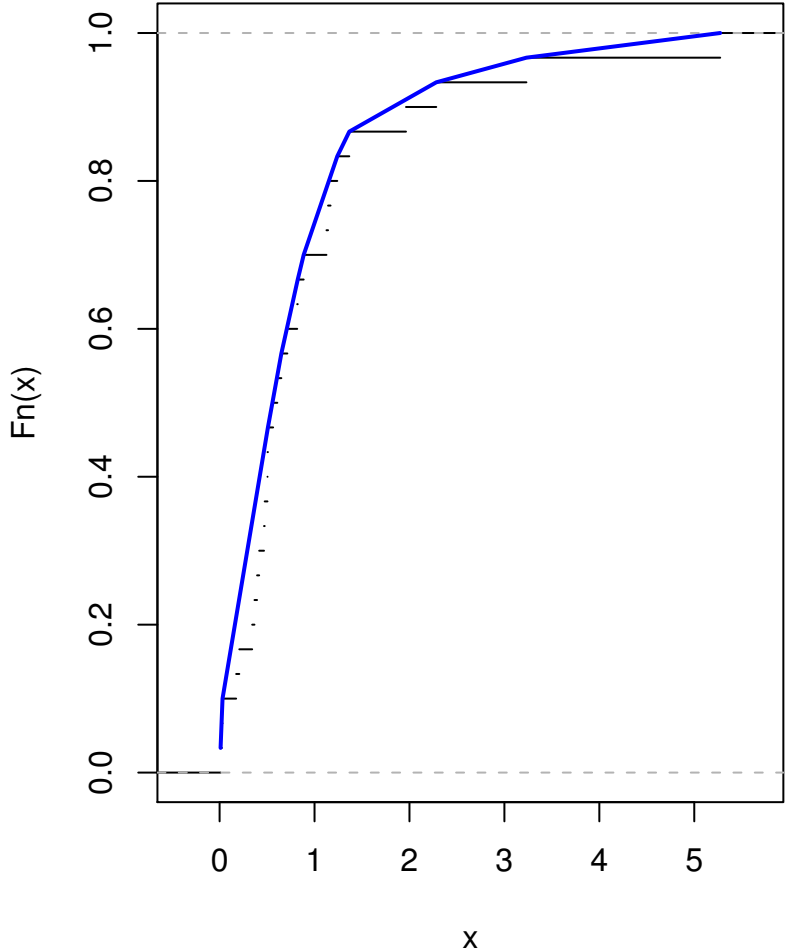
Figure 1
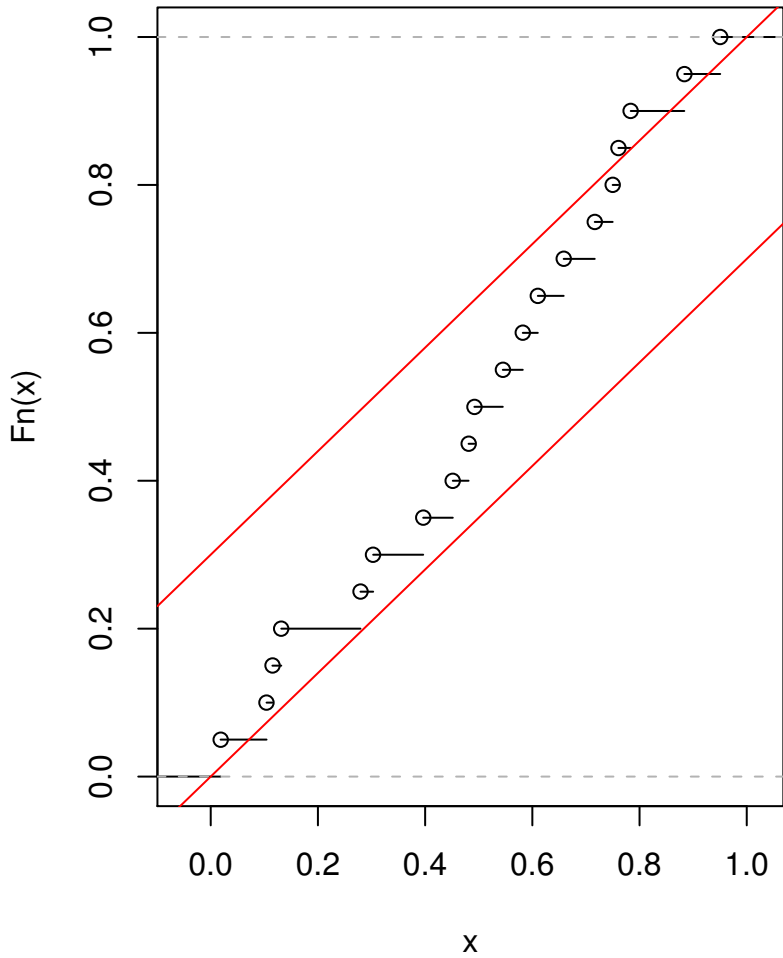
Figure 2

ecdf(pval) ecdf.pval(pval, eta0 = 0.7)
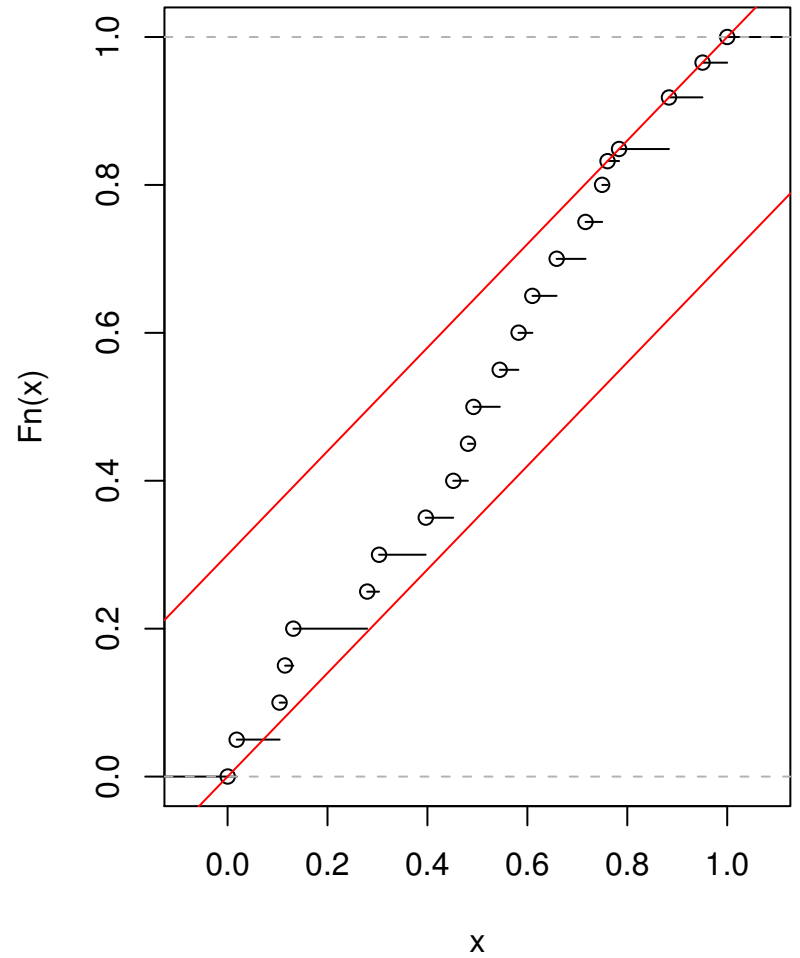
Figure 3

Figure 4
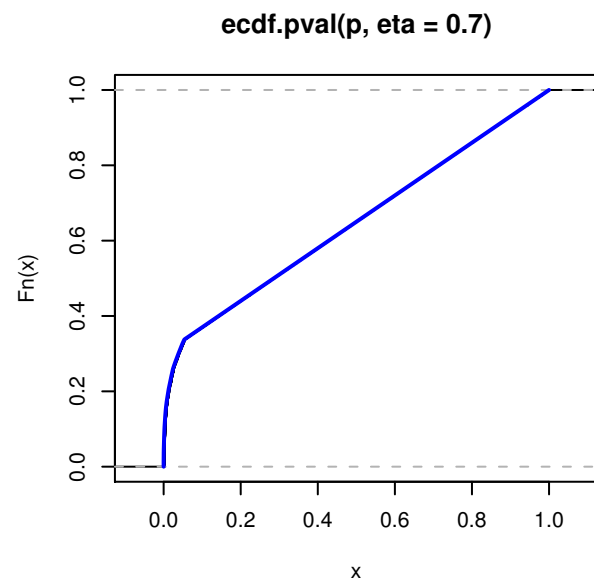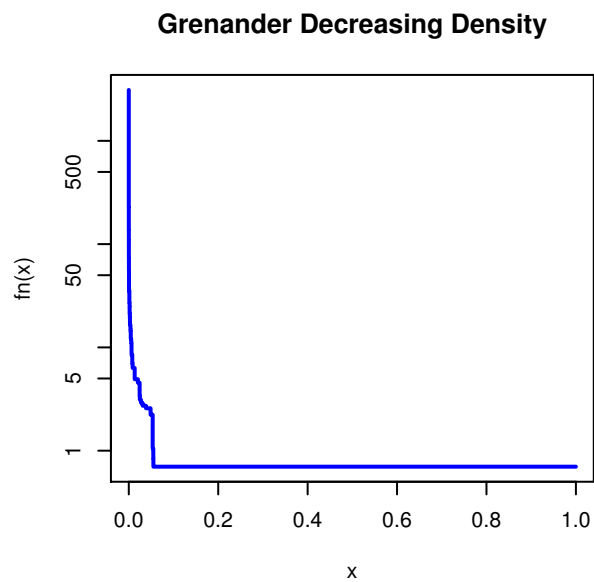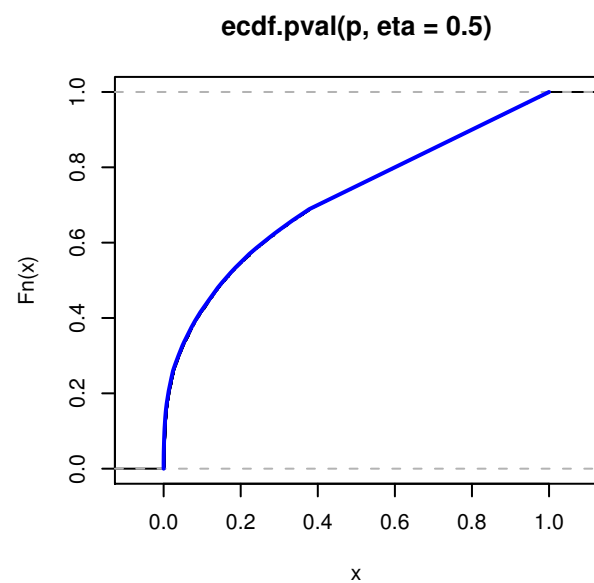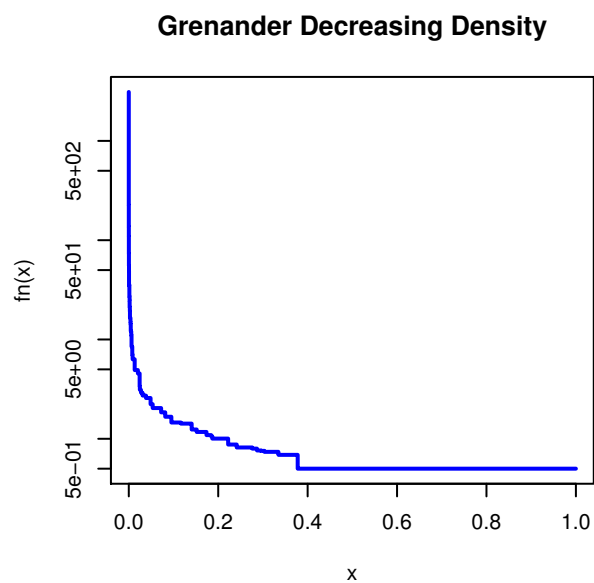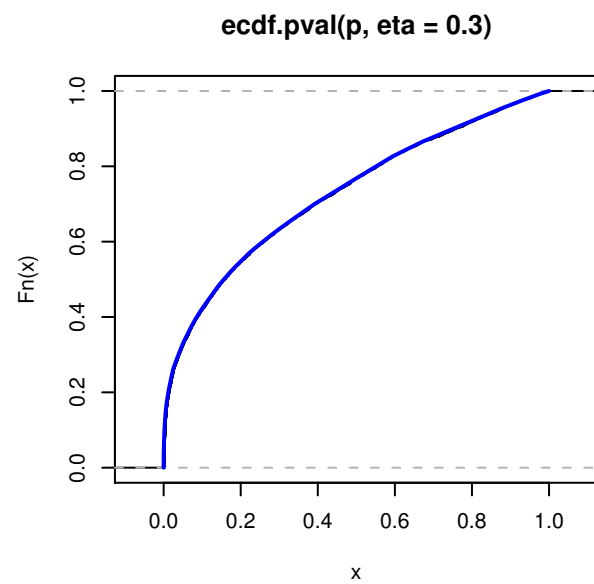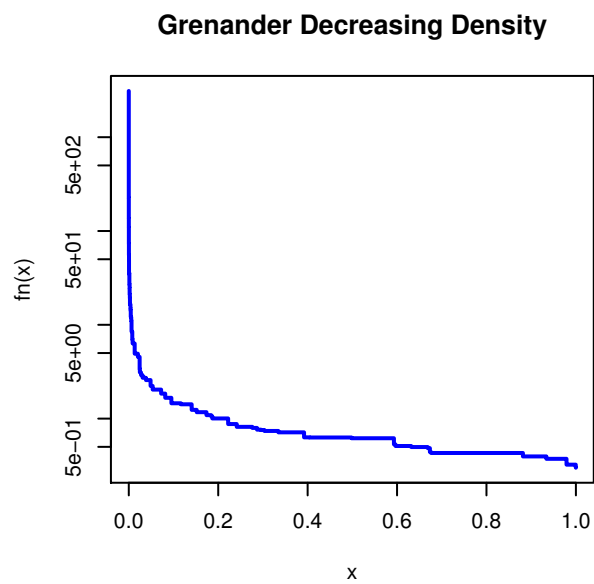
Model 1:
uniform–exponential, a=5
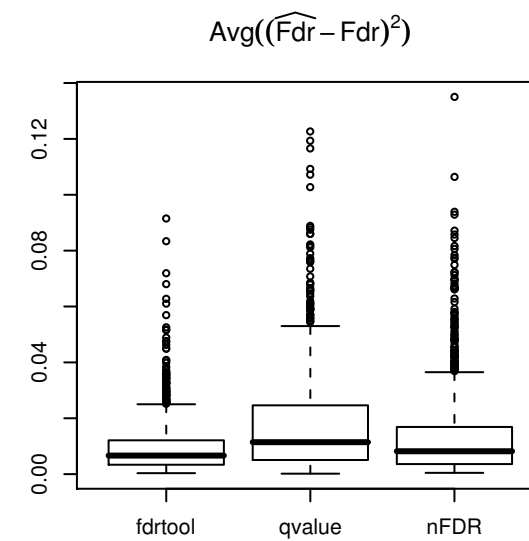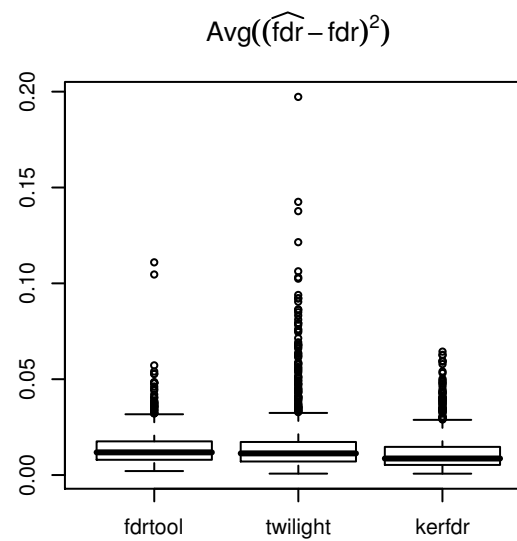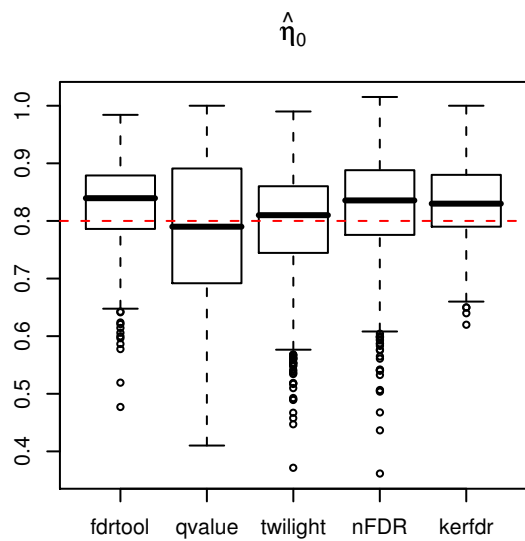
Model 2:
uniform–exponential, a=20
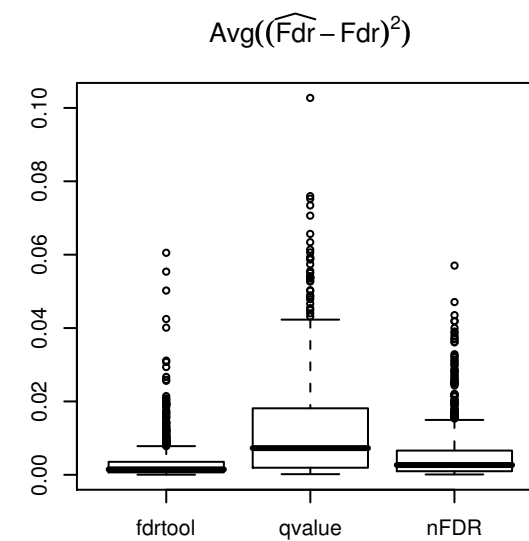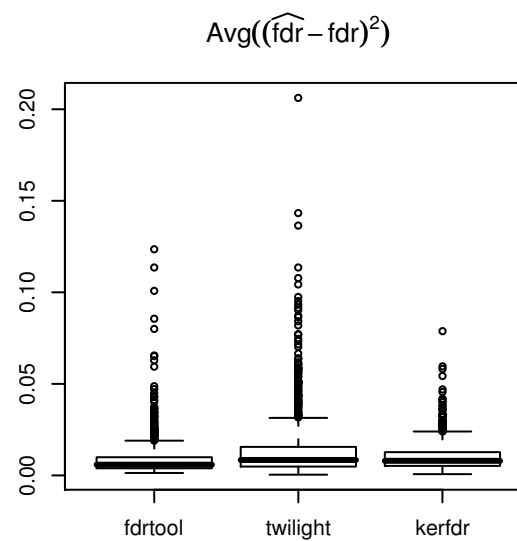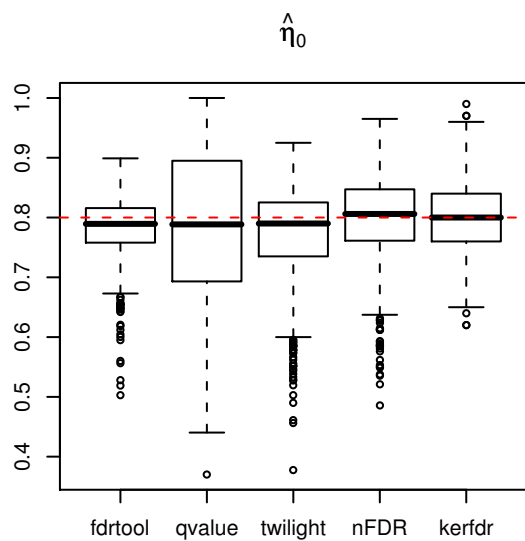
Model 3:
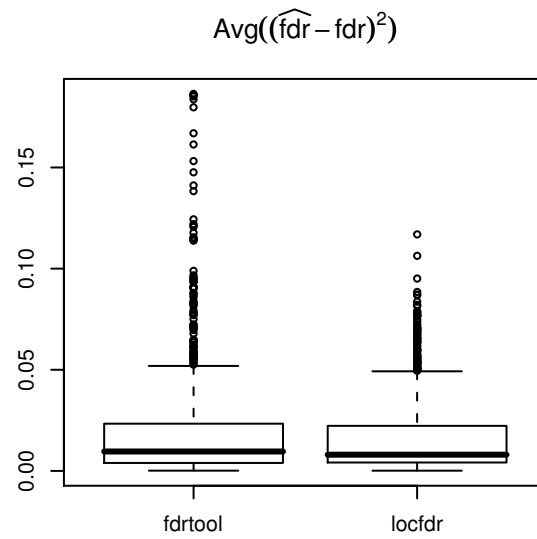uniform–uniform, a=0.2

Figure 5

$\hat{\eta}_0$ $\hat{\sigma}$
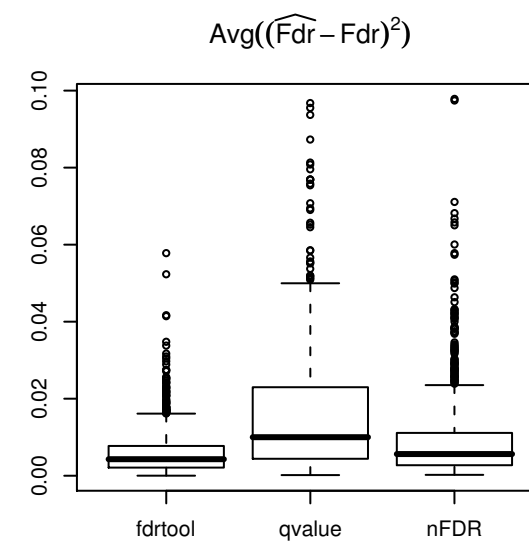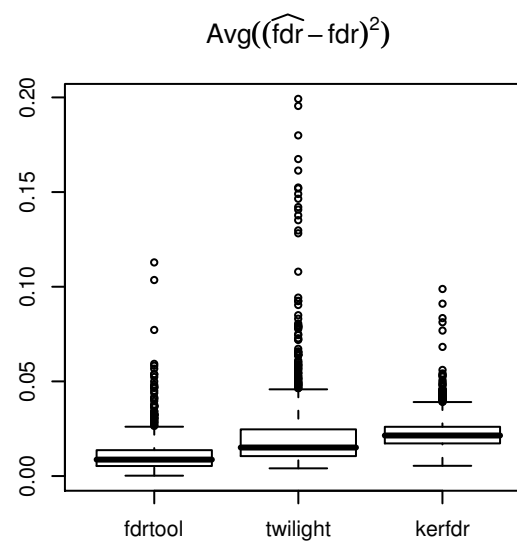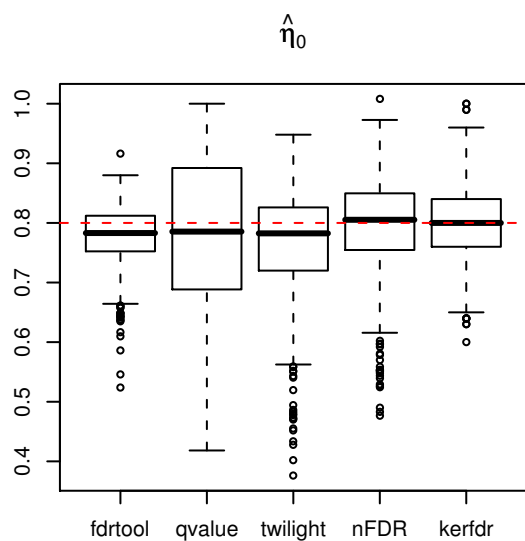
Emprirical Null Parameter Estimates

$\mathrm{Avg}((\widehat{fdr} - fdr)^2)$ $\mathrm{Avg}((\widehat{Fdr} - Fdr)^2)$
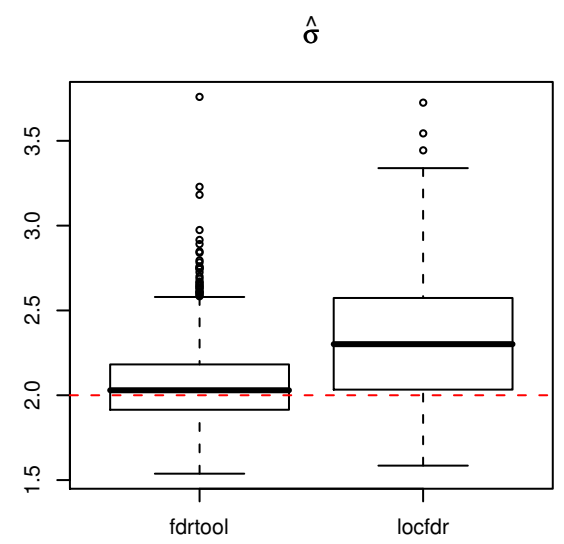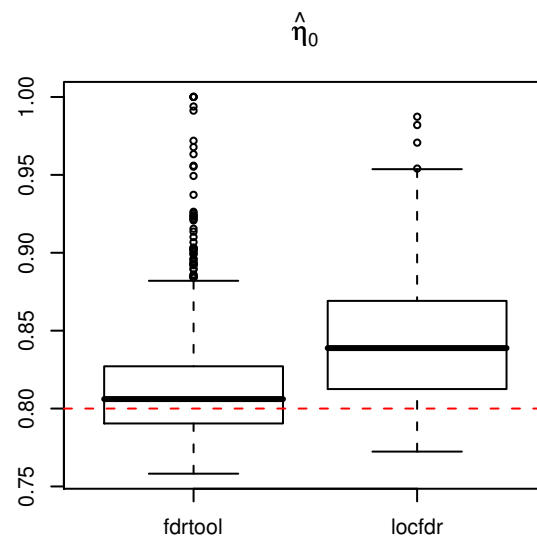
FDR Error
across all z−scores

$\mathrm{Avg}((\widehat{fdr} - fdr)^2)$ $\mathrm{Avg}((\widehat{Fdr} - Fdr)^2)$
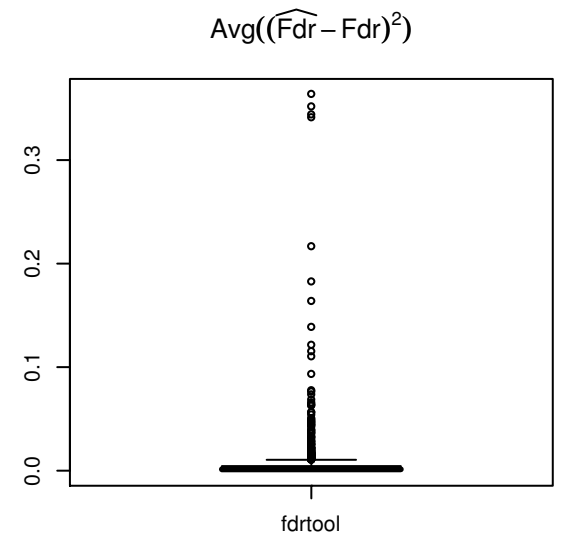
FDR Error
large z−scores only

Figure 6

Figure 7