Monotone false discovery rate

Joong-Ho Won∗1, Johan Lim2, Donghyeon Yu2, Byung Soo Kim3, and Kyunga Kim4

1Korea University
2Seoul National University
3Yonsei University
4Sookmyung Womens University

Abstract

This paper proposes a simple procedure to obtain monotone estimates of both the local and the tail false discovery rates that arise in large-scale multiple testing. The proposed monotonization naturally defines an asymptotically optimal decision rule for controlling the false discovery rate. Monotone false discovery rates also have many attractive features in finite-sample settings. The merit of the proposed procedure is demonstrated with both numerical and microarray data.

Keywords: adaptive decision rule, false discovery rate, empirical Bayes methods, mode matching, isotonic regression.

1 Introduction

The advance of modern high-throughput technologies in many scientific disciplines such as genomics and brain imaging has dramatically increased both the size and the dimension of the data and made data analysis a major challenge. In particular, it is often required to test thousands or millions of hypotheses simultaneously when analyzing large-scale, high-dimensional data. Unlike the case of testing a single hypothesis, type I error in multiple hypothesis testing is not uniquely defined. Traditional approaches, e.g., the family-wise error rate (FWER), are far too conservative and produce many false negatives in high-dimensional settings. For this reason, the concept of false discovery rate (FDR), or the expected proportion of false positives among the declared positives, is introduced and now widely accepted.

The FDR is originally proposed by Benjamini and Hochberg (1995), who develop a step-wise procedure to control the FDR. Storey (2002) proposes to estimate the FDR of a fixed
rejection region and introduces the $q$-value, which is the minimum FDR level to reject the null hypothesis given observed data. Both Benjamini and Hochberg’s and Storey’s procedures assume independence of statistics for hypotheses. Since this assumption does not always hold in practice, Efron has recently introduced an empirical Bayes (EB) procedure based on a two-group mixture model \cite{Efron2004,Efron2007a}. The EB procedure uses the $z$-values instead of the $p$-values and fits them using the two-group mixture model. The EB framework introduces two variants of the FDR: the local FDR, denoted by “fdr”, is the ratio of the null sub-density to the marginal mixture density of the two-group model; the tail FDR, denoted by “Fdr”, is the ratio of the null sub-survival function (tail probability) to the marginal survival function. The EB procedure estimates the null and the marginal mixture distributions from the data. Hence it takes into account the dependence among the test statistics. The estimated null distribution is referred to as the empirical null.

The main theme of this paper is monotonicity in the FDR. Monotonicity is desirable in many settings as it maintains the order of the observed test statistics. In particular, the monotonicity condition for the local FDR implies the SMLR condition of Sun and Cai \cite{Sun2007} (see Section 3.1 for the definition of the SMLR condition). Under this condition, the local FDR yields the optimal oracle decision rule, and a monotone estimate of the local FDR results in a data-driven decision rule that is, under some regularity conditions, asymptotically optimal. Furthermore, a monotone estimate of the local FDR satisfies the SMLR condition in finite-sample settings, which by itself is desirable in practice.

Despite many attractive features of monotonicity, unfortunately, few existing procedures to estimate fdr or Fdr take monotonicity into account. Broberg \cite{Broberg2005} investigates the use of monotone FDR in the setting that the theoretical null distribution of $p$-values is uniform on $[0,1]$. In this setting, monotonicity of fdr (resp. Fdr) is equivalent to that of the marginal density function (resp. the marginal survival function). Monotonicity is enforced by estimating the marginal density function (resp. the marginal survival function) under appropriate constraints, either parametrically or non-parametrically. A similar procedure is employed by Strimmer \cite{Strimmer2008}. For more flexible EB procedures \cite{Efron2007a}, however, one has to estimate both the null and the marginal distributions. We undertake to see how to impose monotonicity in this setting.

We begin with a review of the empirical Bayes theory of false discovery rate. In Section 3, attractive statistical properties of the monotone FDR are discussed. We show that monotonicity in the local FDR is equivalent to that in the likelihood ratio of the components of the two-group mixture model, and implies that of the tail FDR. We propose a procedure that ensures monotonicity in the estimates of the local and the tail FDRs, and that naturally leads to an adaptive decision rule using the monotonized estimates. In Section 4, we conduct a numerical study to demonstrate that the monotonized FDR can improve the performance of the FDR estimates. In Section 5, we illustrate that the proposed procedure can improve real-world data analyses. Section 6 concludes the paper.
2 Empirical Bayes Theory of False Discovery Rates

This section reviews the empirical Bayes theory of false discovery rate inference, largely
developed by Efron (Efron, 2004, 2007a,b).
Suppose we have a collection of $N$ hypotheses and their corresponding “summarizing
statistics” $T_1, \ldots, T_N$. Assume that the $T_i$s have a common marginal distribution whose
density is of the two-group mixture form:
\[
f(t) = p_0 f_0(t) + p_1 f_1(t),
\]
where $f_0(t)$ and $f_1(t)$ are the null and the non-null densities, respectively; $p_0$ is the proportion
of the null group, and $p_1 = 1 - p_0$. We define the null sub-density as $p_0 f_0(t)$. The local
false discovery rate (denoted by $fdr$) and the (right-) tail FDR (denoted by $Fdr$) at $t$ are,
respectively, defined as
\[
fdr(t) = \frac{p_0 f_0(t)}{f(t)} \quad \text{and} \quad Fdr(t) = \frac{p_0 S_0(t)}{p_0 S_0(t) + p_1 S_1(t)},
\]
where $S_0(t)$ and $S_1(t)$ are the survival functions of the null and the non-null groups, respectively.
Knowledge of the null density $f_0(t)$ plays a crucial role in the inference regarding $fdr$ and $Fdr$. The null distribution of the summarizing statistics for single hypothesis testing is often
known theoretically, e.g., standard normal, Student’s $t$, or chi-square. However, in multiple
hypothesis testing, the observed summarizing statistics often do not follow the theoretical
null distribution. This phenomenon may be due to failed assumptions, unobserved covariates,
correlations among the samples or among the summarizing statistics (Efron, 2007b).
To remedy this problem, several authors advocate a family of empirical Bayes procedures,
referred to as the empirical null method (Efron, 2007a,b; Schwartzman, 2008). This method
estimates the null distribution from the data itself. For $N$ sufficiently large, the components
of the mixture density (1) can be estimated under a certain set of assumptions. These
assumptions include that $f_0(t)$ is unimodal, and that the most of the probability mass around
the peak of $f(t)$ is due to the null sub-density $p_0 f_0(t)$. Therefore, a reliable estimation of
$f_0(t)$ and $p_0$ is very important for accurate inference of the FDRs discussed above.
To estimate $f(t)$, $f_0(t)$, and $p_0$, Efron (2007b) proposes two methods, namely “central
matching” and “MLE fitting”. First, central matching is a two-step procedure. At step 1,
the mixture density $f(t)$ is modeled as a semi-parametric exponential family, e.g.,
\[
f(t) = c_\beta \exp \{ \sum_{j=1}^7 \beta_j t^j \},
\]
where $c_\beta$ is a normalization constant. Subsequently the $N$ test statistics
are binned into $K$ bins with equal width $\Delta$ centered at $t_1, t_2, \ldots, t_K$. Let $y_k$ be the count
in bin $k$. Then the parameters $\{\beta_j\}$ are fitted to $\{y_k\}$ using Lindsey’s method (Lindsey,
1974). At step 2, $f_0(t)$ is fit to the estimated $f(t)$ around $t = 0$. Assuming $f_0(t)$ is a
normal density, the parameters (mean and variance) for $f_0(t)$ are estimated by least squares.
Second, MLE fitting undertakes maximum likelihood estimation, in which it is assumed that
the non-null density is only supported outside some known interval $[t_{\min}, t_{\max}]$, i.e.,
$f_1(t) = 0$ for $t \in [t_{\min}, t_{\max}]$, and the null density is normal with unknown mean and variance. The
likelihood function of the $N$ summarizing statistics is a product of a binomial and a truncated normal likelihoods. Then the parameters, i.e., $p_0$ and the mean and the variance of the null, are estimated by maximizing the product likelihood.

Central matching has been further generalized to general exponential families by Schwartzman (2008) (“mode matching”). Assuming that the null density is taken from an exponential family $f_0(t) = a_0(t) \exp(x(t)^T \eta - \psi(\eta))$, and the counts $y_k$ within $[t_{\min}, t_{\max}]$ are independent Poisson variables with mean $\lambda_k \approx N \Delta p_0 f_0(t_k)$, the following Poisson regression model is obtained:

$$\log(\lambda) = X \eta^+ + h,$$

where $\lambda = (\lambda_1, \ldots, \lambda_K)^T$; $\eta^+ = (C, \eta)^T$ with $C = \log p_0 - \psi(\eta)$; $X$ is the design matrix with rows $(1, x(t_k)^T)^T$, $k = 1, \ldots, K$; and $h = (h_1, \ldots, h_K)$ is a known offset vector with $h_k = \log(N \Delta a_0(t_k))$. Solving (3) provides an estimate vector $(\hat{C}, \hat{\eta})^T$, from which the proportion of the null group $\hat{p}_0 = \exp(\hat{C} + \psi(\hat{\eta}))$ is reconstructed. Then the estimates of the fdr and the Fdr at the bin centers $\{t_k\}$ are evaluated by

$$\hat{\text{fdr}}(t_k) = \frac{p_0 \hat{f}_0(t_k)}{\hat{f}(t_k)} = \frac{\hat{y}_k}{y_k} \quad \text{and} \quad \hat{\text{Fdr}}(t_k) = \frac{(1/2)\hat{y}_k + \sum_{j=k+1}^K \hat{y}_j}{(1/2)\hat{y}_k + \sum_{j=k+1}^K y_j},$$

where $(\hat{y}_1, \ldots, \hat{y}_K)^T = \hat{y} = \exp(X \hat{\eta}^+ + h)$ is the vector of the expected frequencies of the bins. Equivalently, in vector form, they are written as

$$\log \hat{\text{fdr}} = \log \hat{y} - \log y \quad \text{and} \quad \log \hat{\text{Fdr}} = \log (S \hat{y}) - \log (Sy),$$

where $S$ is an upper triangular matrix with entries $1/2$ on the diagonal and 1 above the diagonal.

### 3 Monotone False Discovery Rate

In this section, we first examine attractive properties of the monotone FDR (local FDR and tail FDR). We then propose a procedure to monotonize the estimates of $\text{fdr}(t)$ and $\text{Fdr}(t)$, and adaptive optimal procedure using the monotonized estimates. In the remainder of the section, we assume that $\text{fdr}(t)$ is monotonically decreasing.

#### 3.1 Properties of monotone FDR

**Monotone local FDR implies monotone likelihood ratio** Recall that

$$\text{fdr}(t) = \frac{p_0 f_0(t)}{p_0 f_0(t) + p_1 f_1(t)} = \frac{p_0}{p_0 + p_1 f_1(t)/f_0(t)}.$$ 

This shows that monotone decrease of $\text{fdr}(t)$ is equivalent to monotone increase of the likelihood ratio $f_1(t)/f_0(t)$. This equivalence in turn defines a stochastic ordering between the null and the alternative densities: the alternative density $f_1$ is said to be stochastically larger than the null density $f_0$ if the likelihood ratio is monotonically increasing (Robertson et al., 1988; Lim and Won, 2012). A similar statement can be made for the tail FDR.
Monotone local FDR and the SMLR condition  For a random variable $T$ that has the identical distribution to the common marginal distribution of $T_1, \ldots, T_N$, write the marginal density of $V = \text{fdr}(T)$ as $p_0 g_0(v) + p_1 g_1(v)$. Here $g_0$ and $g_1$ are the conditional densities of $V$ under the null and the non-null, respectively. Since fdr$(t)$ is monotone decreasing in $t$, fdr$^{-1}(t)$ is well-defined and the following likelihood ratio

$$
\frac{g_1(v)}{g_0(v)} = \frac{f_1(fdr^{-1}(v))/(fdr^{-1}(v))'}{f_0(fdr^{-1}(v))/(fdr^{-1}(v))'} = \frac{f_1(fdr^{-1}(v))}{f_0(fdr^{-1}(v))}
$$

is decreasing in $v$. Hence the oracle statistic fdr$(T)$ has a monotone likelihood ratio. This is precisely the SMLR condition of Sun and Cai (2007). Note that if a statistic $T(T)$ satisfies the SMLR condition, the decision rule $I\{T(T) < c\}$ has many attractive features for multiple testing problems:

**Proposition 1.** ([Sun and Cai, 2007, Proposition 1]). When $N$ summarizing statistics $T_1, T_2, \ldots, T_N$ follows the two-group mixture model (1), if a statistic $T(T_i)$ satisfies the SMLR condition, then applying the decision rule $I\{T(T_i) < c\}$ for $i = 1, \ldots, N$ implies

1. $\Pr(\text{non-null} | T(T_i) < c)$ is monotonically decreasing in threshold $c$,

2. $mFDR$ is monotonically increasing in $c$ and the expected number of rejections $r$, and

3. $mFNR$ is monotonically decreasing in $c$, $r$, and $mFDR$,

where mFDR is the marginal false discovery rate, or $\Pr(T(T_i) < c, \text{null}) / \Pr(T(T_i) < c)$, and mFNR is the marginal false non-discovery rate, or $\Pr(T(T_i) > c, \text{non-null}) / \Pr(T(T_i) > c)$.

Optimality of the monotone local FDR  Not only that it has many good properties as a statistic for multiple testing, the monotone local FDR is optimal:

**Theorem 1.** If the local FDR in (2) is monotonically decreasing, then for any given mFDR level $\alpha$ in a multiple testing problem on the summarizing statistics $T_1, T_2, \ldots, T_N$, there exists a unique $c(\alpha)$ such that the decision rule $I\{\text{fdr}(T_i) < c(\alpha)\}$ has an mFDR not greater than $\alpha$ and the smallest mFNR among all decision rules of the form $I\{T(T_i) < c\}$, where $T$ satisfies the SMLR condition and $c$ can be any constant.

**Proof.** This can be easily proven by using Theorems 1 and 2 of Sun and Cai (2007), and the above result that the SMLR condition holds for the monotone local FDR.

Monotonicity of the local, tail, and marginal FDRs  If fdr$(t)$ is monotonically decreasing, then the mFDR of the decision rule $I\{\text{fdr}(T) < c\}$ is written as

$$
\frac{\Pr(\text{fdr}(T) < c, \text{ null})}{\Pr(\text{fdr}(T) < c)} = \frac{p_0 S_0(\text{fdr}^{-1}(c))}{p_0 S_0(\text{fdr}^{-1}(c)) + p_1 S_1(\text{fdr}^{-1}(c))} = \text{Fdr}(\text{fdr}^{-1}(c)).
$$
By Proposition 1, mFDR is monotonically increasing in $c$. Hence $\text{Fdr}(t)$ is monotonically decreasing in $t = \text{fdr}^{-1}(c)$. Furthermore, the tail FDR can be controlled by controlling the local FDR:

**Proposition 2.** Assume $f_0(t)$ and $f_1(t)$ are continuous and positive for every $t$. If $\text{fdr}(t)$ is monotonically decreasing in $t \in \mathbb{R}_+$, then for every $\alpha \in (0, 1)$, we have

$$\{ t : \text{fdr}(t) \leq \alpha \} \subset \{ t : \text{Fdr}(t) \leq \alpha \}.$$

**Proof.** Let $\alpha$ be an arbitrary number between 0 and 1, and let $t_\alpha$ be the unique root of the equation $\text{fdr}(t) = \alpha$. Then, for $t \geq t_\alpha$, $f_1(t)/f_0(t) \leq f_1(t_\alpha)/f_0(t_\alpha) = (p_0/\alpha - p_0)/p_1$. Now, with the definition of $\text{fdr}(t)$, a simple algebra shows

$$\frac{S_1(t_\alpha)}{S_0(t_\alpha)} = \frac{1}{S_0(t_\alpha)} \int_{t_\alpha}^{\infty} f_0(s) \left( \frac{f_1(s)}{f_0(s)} \right) ds \geq \frac{1}{S_0(t_\alpha)} \int_{t_\alpha}^{\infty} f_0(s) ds \cdot \left( \frac{p_0}{\alpha} - p_0 \right) \frac{1}{p_1} = \left( \frac{p_0}{\alpha} - p_0 \right) \frac{1}{p_1},$$

hence

$$\text{Fdr}(t_\alpha) = \frac{p_0}{p_0 + p_1 S_1(t_\alpha)/S_0(t_\alpha)} \leq \alpha.$$

---

**Estimated local FDR statistic** As a final note, recall that $\text{fdr}(T)$, is not a *bona fide* statistic unless $f_0$, $f_1$, and $p_0$ are known *a priori*. As a solely data-driven, hence *bona fide*, statistic, we may consider an estimator $\hat{\text{fdr}}(T)$ of $\text{fdr}(T)$, using the methods in Section 2. However, the resulting finite-sample estimator of the local FDR is not necessarily monotone, even if the true local FDR is. Hence it is desirable to incorporate monotonicity in the estimation procedure.

Post-hoc monotonization of the local FDR estimates in the next section is attractive in the following sense. If the true local FDR is monotone, then the monotonized local FDR satisfies the SMLR condition and yields a decision rule that enjoys the good properties listed in Proposition 1. This is readily seen by plugging in a monotone estimator $\hat{\text{fdr}}(\cdot)$ in place of the true (monotone) $\text{fdr}(\cdot)$ in (4). Furthermore, if the non-monotonized estimate is consistent, then monotonization preserves consistency while reducing variance.

### 3.2 Estimation by monotonization

We propose to modify the FDR estimates by imposing a monotone ordering ("isotonization") among them. Suppose the mode matching method by Schwartzman (2008) is employed to estimate the local FDR and the tail FDR. Using the delta method, the variance-covariance matrices of $\log \hat{\text{fdr}}$ and $\log \hat{\text{Fdr}}$ are computed as follows. Let $X$ be the design matrix in (3), and $W$ be the diagonal matrix made of the vector $w = (w_1, w_2, \ldots, w_K)$, where $w_k$ is equal to 1 or 0 according to whether $t_k$ is in the null region $[t_{\min}, t_{\max}]$ for the Poisson regression.
Set $\tilde{V} = \text{diag}(\tilde{y})$, $\tilde{V}_N = \tilde{V} - \tilde{y}\tilde{y}^T/N$, and $D_y = X(X^T W \tilde{V} X)^{-1} X^T W$. Then the desired variance-covariance matrices are given as

$$\tilde{\text{cov}}(\log \hat{\text{fdr}}) = A \tilde{V}_N A^T \quad \text{and} \quad \tilde{\text{cov}}(\log \hat{\text{Fdr}}) = B \tilde{V}_N B^T,$$

where $A = D_y - V^{-1}$ and $B = \tilde{U}^{-1} S \tilde{V}^{-1} D_y - U^{-1}$ with $U = \text{diag}(S \tilde{y})$ and $\tilde{U} = \text{diag}(S \tilde{y})$ (Schwartzman, 2008).

We proceed to adjust the initial estimate $\hat{\text{fdr}}$ by solving the quadratic program (QP):

$$\text{minimize} \quad (z - \log \hat{\text{fdr}})^T \tilde{\text{cov}}(\log \hat{\text{fdr}})^{-1} (z - \log \hat{\text{fdr}})$$

subject to $z_1 \leq z_2 \leq \cdots \leq z_K$. This QP is a convex optimization problem that can be efficiently solved using existing software packages, e.g., quadprog R package.

If $K$ is large, we suggest to solve a simplified version of (5):

$$\text{minimize} \quad (z - \log \hat{\text{fdr}})^T \text{diag}\left\{\tilde{\text{cov}}(\log \hat{\text{fdr}})^{-1}\right\} (z - \log \hat{\text{fdr}})$$

subject to $z_1 \leq z_2 \leq \cdots \leq z_K$. This is a generalized isotonic regression problem and can be solved using the pool-adjacent-violator (PAVA) algorithm of Robertson et al. (1988). A similar procedure can be applied to monotonize $\hat{\text{Fdr}}$, estimates of Fdr.

### 3.3 Adaptive Decision Rule with Monotonized Estimate

Suppose we have obtained monotonized estimates $\hat{\text{fdr}}_{iso}(t_k)$ and $\hat{\text{Fdr}}_{iso}(t_k)$ of the local and the tail FDRs at $t = t_k$, respectively. Let $\hat{\text{fdr}}_{iso}(k)$ and $H_{(k)}$ be the $k$th largest value and its corresponding null hypothesis. Following Sun and Cai (2007), we propose a decision rule that is step-up and rejects all hypotheses $H_{(k)}$, $k = 1, 2, \ldots, u$, where

$$u = \max \left\{ j \mid (1/j) \sum_{k=1}^{j} \hat{\text{fdr}}_{iso}(k) \leq \alpha \right\} \quad \text{and} \quad \hat{\text{Fdr}}_{iso}(j) \leq \alpha$$

for the local FDR and the tail FDR, respectively.

### 4 Numerical Study

In this section, we compare the performance of the monotonized FDR estimators in Section 3.2 to the unconstrained estimators numerically. The same numerical scheme as in Schwartzman (2008, Section 4) is used for this study.
Consider two scenarios to generate sets of summarizing statistics from the two-group mixture model (1): $T_i$s are independent random variables from the following mixture models

\[
T_i \sim \begin{cases} 
  f_0 = N(0.2, 1.2^2) & \text{w.p. } p_0, \\
  f_1 = N(3, 1.2^2) & \text{w.p. } 1 - p_0,
\end{cases}
\]

or

\[
T_i \sim \begin{cases} 
  f_0 = 0.8\chi^2(3) & \text{w.p. } p_0, \\
  f_1 = \chi^2(3, 3) & \text{w.p. } 1 - p_0,
\end{cases}
\]

where $N(\mu, \sigma^2)$ denotes the normal density with mean $\mu$ and variance $\sigma^2$, $a\chi^2(\nu)$ denotes the scaled chi-square distribution with $\nu$ degrees of freedom, and $\chi^2(\nu, \delta)$ denotes the non-central chi-square distribution with the non-centrality parameter $\delta$. In the study, we assume $p_0 = 0.9$ and generate 100 data sets for each case. The fitting interval to estimate the null distribution is set to be $[0.2 - t_0, 0.2 + t_0]$ for the normal case and $[0, t_0]$ for the $\chi^2$ case. We fix the bin width $\Delta = 0.1$ and $t_0 = 1$ and apply the mode matching method and our proposal to estimate $\text{fdr}(t)$ and $\text{Fdr}(t)$, with and without the monotonicity constraint.

Figure 1 plots the average of 100 fdr estimates and its 95% confidence intervals of each method. Figure 1(a) and (c) indicates that the monotonized fdr estimates have smaller variance than their unconstrained counterparts. In particular, the unconstrained estimates for the chi-square case are quite volatile for large $t$ values, even after taking an average of the 100 estimates. This volatility is substantially reduced after monotonization. Furthermore, the confidence interval for the monotinized estimates is much narrower than the unconstrained one.

The Fdr estimates are on average quite smooth and monotone as compared to the fdr estimates. In Figures 1(b) and (d) the averages of 100 non-monotonized and monotonized Fdr estimates are almost equal. Paying attention to the individual data set, however, 4 out of 100 data sets result in non-monotone Fdr estimates in the normal case; 17 out of 100 result in non-monotone estimates in the chi-square case. If we limit our attention to these cases that do have non-monotone Fdr estimates, we observe that the isotonization step improves accuracy (Figure 2).

5 Examples

In this section, we illustrate the merit of monotonization using the leukemia data by Golub et al. (1999), available from http://www.broadinstitute.org/cancer/pub/all_aml/. This data set records the expression levels of patients with one of the two types of leukemia, acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). The data set consists of two parts: training and test. The training data set comprises of 38 arrays (ALL, 27; AML, 11). The test data set has 34 arrays (ALL, 20; AML, 14). In our analysis, we only used the training set, as in Broberg (2005), to find differentially expressed genes (DEGs) between ALL and AML. Preprocessing was conducted according to the prescription due to Dudoit et al. (2002). The preprocessed data set was summarized as a $38 \times 3571$ matrix. (This data set is available from R package multtest).

We applied the mode matching procedure (Schwartzman, 2008) and monotonized the estimated fdrs and Fdrs using the method of Section 3.2 to find DEGs. For $g = 1, 2, \ldots, G = 3571$, we computed two-sample $t$-statistics (with equal variance) $t_g$ as summarizing statistics,
and transformed them to z-values $z_g = \Phi^{-1}(F_{36}(t_g))$, where $F_{36}(t)$ is the cumulative distribution function of the $t$-distribution with 36 degrees of freedom. In applying the mode matching procedure, we chose the bin size $\Delta = 0.05$ and the null region $[-1.2, 1.2]$ to estimate the empirical null distribution. We found the mode-matched estimates of the fdr's frequently violate monotonicity, but the estimates of Fdr's did not need additional monotonization; fdr's are non-smooth and not monotone due to scarcity of observations in both tails. We only monotonized the fdr's outside the null region, i.e., those in the region $(-\infty, -1.2) \cup [1.2, \infty)$. Figure 5 depicts the estimates of the local FDR and their monotonization.

We then applied the adaptive decision rule of Section 3.3 to the non-monotonized and the monotonized estimates of the local FDR to declare DEGs. We control the marginal FDR level at $\alpha = 0.05, 0.1$, and 0.15. At level $\alpha = 0.05$, we found 68 DEGs using the non-monotonized fdr estimates (denoted by unadj) and 40 DEGs by using their monotonized modifications (denoted by iso). Using HuGE Navigator version 2.0 databases (Yu et al., 2009).
Figure 2: Fdr for chi-square distribution. The plot of averages of non-monotone Fdr estimates and their monotonized estimates. Only the right-hand side of the vertical line is monotonized.

Figure 3: Monotonized local FDR estimates of leukemia data in Golub et al. (1999). “o” plots un-monotonized estimates, and the solid line represents the monotonized local FDR estimates.

In 2008, we investigated the biological relevance of the DEGs found by seeking AML/ALL-related genes among those genes. Our data set has 2625 unique genes (from the 3571 probes), among which the number of AML/ALL-related genes reported by HuGE navigator was 130 (4.4% of the 2625 genes). Our monotonization removed 32 genes from the 68
DEGs that unadj found, and introduced 4 new genes, one of which was AML/ALL-related; only 2 of the 32 DEGs removed were AML/ALL-related genes. In short, the percentage of AML/ALL-related genes in detected DEGs increased from 8.82% to 12.5% by taking into account monotonicity in the local FDR. This observation indicates that the isotonization can reduce the number of false discoveries. This observation is still valid for levels $\alpha = 0.1$ and 0.15, although the improvement due to isotonization becomes smaller as the level increases. These results are summarized in Table 1.

Table 1: Biological relevance of detected DEGs

| $\alpha$ | FDR      | # of DEGs | # of AML/ALL-related | % of AML/ALL-related |
|---------|----------|-----------|---------------------|---------------------|
| 0.05    | fdr (iso) | 40        | 5                   | 12.50%              |
|         | fdr (unadj) | 68        | 6                   | 8.82%               |
| 0.1     | fdr (iso) | 125       | 9                   | 7.20%               |
|         | fdr (unadj) | 177       | 11                  | 6.21%               |
| 0.15    | fdr (iso) | 232       | 14                  | 6.03%               |
|         | fdr (unadj) | 362       | 21                  | 5.80%               |

As another example, we analyzed the adenocarcinoma data set in Notterman et al. (2001). The data consist of 18 subjects for each of which 6579 gene expressions in the adenocarcinoma and normal colon samples are obtained and paired. The results are reported in the supplementary material, and again indicate that monotonization can reduce the number of false discoveries.

6 Conclusion

We have considered monotonicity in the FDR and proposed an estimation procedure thereof. The proposed procedure is a simple modification of the empirical Bayes estimator using generalized isotonic regression. The presented numerical study shows that imposing monotonicity improves the estimation in both bias and variance. Through real-world data sets, it is demonstrated that the proposed monotone FDR procedure can reduce the number of false discoveries.

Monotonicity in the FDR has several attractive features: monotone local FDR implies optimality in controlling the tail FDR, and monotonized estimates perform better than their non-monotonized counterparts in practice. The latter may be due to that imposing smoothness (via monotonization) improves estimation as in many non-parametric regression problems.
References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society – Series B, 57*, 289–300.

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

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*, 77–87.

Efron, B. (2004). Large scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association, 99*, 96–104.

Efron, B. (2007a). Correlation and large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association, 102*, 93–103.

Efron, B. (2007b). Size, power and false discovery rates. *The Annals of Statistics, 35*, 1351–1377.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science, 286*, 531–537.

Lim, J. and Won, J.-H. (2012). ROC convex hull and nonparametric maximum likelihood estimation. *Machine Learning, 88*, 433–444.

Lindsey, J.K. (1974). Construction and comparison of statistical models. *Journal of the Royal Statistical Society - Series B, 36*, 418–425.

Notterman, D.A., Alon, U., Sierek, A.J., and Levine, A.J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research, 61*, 3124–3130.

Robertson, T., Wright, F., and Dykstra, R. (1988), *Order Restricted Statistical Inference*. Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons.

Schwartzman, A. (2008). Empirical null and false discovery rate inference for exponential families. *The Annals of Applied Statistics, 35*, 1351–1377.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics, 9*, 303.
Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society – Series B*, 64, 479–498.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102, 901–912.

Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., Khoury, M.J. (2008). A navigator for human genome epidemiology. *Nature Genetics*, 40, 124–125.
Supplementary material for
“Monotone false discovery rate”

Joong-Ho Won*1, Johan Lim2, Donghyeon Yu2, Byung Soo Kim3, and Kyunga Kim4

1Korea University
2Seoul National University
3Yonsei University
4Sookmyung Womens University

Adenocarcinoma Example

In this document, we provide an additional data example to show an advantage of the proposed monotone estimates of the local FDR. We analyzed the adenocarcinoma data set in Notterman et al. (2001), which is available from http://genomics-pubs.princeton.edu/oncology. The data set consists of 18 arrays of adenocarcinoma samples and another 18 arrays of their paired normal colon samples. Each array records expression levels of 7,457 genes. In the array, expression levels of some genes were repeatedly measured and they are replaced with their averages. After this, 6,579 genes remained in each array.

As in the Example section (Section 5 of main paper), to control the marginal FDR, we apply the adaptive optimal procedure by Sun and Cai (2007) with non-monotonized estimates and monotonized estimates of the local FDR. We controlled the marginal FDR at 0.01, 0.025 and 0.05.

*E-mail: wonj@korea.ac.kr.
We found 61 DEGs by the adaptive procedure with the non-monotonized fdr estimates (denoted by \texttt{unadj}) at 0.01 and 50 DEGs by using their monotonized modifications (denoted by \texttt{iso}) with the same procedure. Using HuGE Navigator version 2.0 databases (Yu et al., 2008), we investigated the biological relevance of the DEGs found by seeking adenocarcinoma neoplasms-related genes among those genes. Our data set has 6,597 unique genes out of 7,457 observed genes, among which the number of adenocarcinoma neoplasms-related genes is 312 (4.73% of 6,597 genes). Our monotonized estimates removed 11 genes from the 61 DEGs found by non-monotonized estimates of the local FDR. All 11 removed genes are not in the list of adenocarcinoma neoplasms-related genes. In short, the percentage of adenocarcinoma neoplasms-related genes in detected DEGs increased from 9.84% to 12% by taking into account the monotonicity of the local FDR. This shows again that the proposed isotonization can reduce the number of false discoveries. We made similar observation at level 0.025 and 0.05 as shown in Table 1. The estimates of local FDR are plotted in Figure 1.

| $\alpha$  | FDR         | # of DEGs | # of Adenocarcinoma neoplasms-related | % of Adenocarcinoma neoplasms-related |
|-----------|-------------|-----------|--------------------------------------|--------------------------------------|
| 0.01      | fdr(iso)    | 50        | 6                                    | 12%                                  |
|           | fdr(unadj)  | 61        | 6                                    | 9.84%                                |
| 0.025     | fdr(iso)    | 101       | 10                                   | 9.90%                                |
|           | fdr(unadj)  | 115       | 11                                   | 9.57%                                |
| 0.05      | fdr(iso)    | 182       | 16                                   | 8.79%                                |
|           | fdr(unadj)  | 215       | 18                                   | 8.37%                                |
Figure 1: The transformed z-statistics are from $z_i = \Phi(F_{34}(t_i))^{-1}$, where $F_{34}$ is the CDF of $t$-distribution with degree of freedom 34 and $\Phi$ is a standard normal CDF. The estimated local FDR is obtained by Schwartzman’s method (Schwartzman, 2008).
References

Notterman, D.A., Alon, U., Sierek, A.J., and Levine, A.J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research*, **61**, 3124–3130.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, **102**(479), 901–912.

Schwartzman, A. (2008). Empirical null and false discovery rate inference for exponential families. *The Annals of Applied Statistics*, **35**(4), 1351–1377.

Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., Khoury, M.J. (2008). A navigator for human genome epidemiology. *Nature Genetics*, **40**, 124–125.