A Semiparametric Approach to Empirical Bayes Estimation of Discrete False Discovery Rates Using Kernels

D. B. Dayta[1] S. M. Escalante[2]

[1]School of Statistics
University of the Philippines, Diliman

[2]*Graduate Student*
Aboitiz School of Innovation, Technology, & Entrepreneurship
Asia Institute of Management

IASC-ARS Interim Conference, December 2022

Agenda

1. Introduction
2. The Multiple Hypothesis Problem
3. Using Discrete Kernels
4. Simulation Results
5. Application to Hedenfalk [2001]

Introduction

Contemporary statistical computing sees a fertile ground of problems in genomics, where inference can be done simultaneously on thousands of genes.

Efron et al. [2001] took the idea of False Discovery Rates and approached it from a Bayesian perspective. Also introduced the local false discovery rate (hereafter referred to as the lowercase 'fdr'), quantifying the probability that a given null hypothesis is true given the observed value of its test statistic sense.

Robin et al. [2007] and Guedj et al. [2009], this mixture model comes center field in a novel estimation procedure for the local false discovery rate involving kernel functions.

The Multiple Hypothesis Problem

Consider testing m null hypotheses simultaneously. In the following table, m_0 denotes the number of true null hypotheses, of which V of them are falsely rejected.

Table: The multiple hypothesis testing problem

		Decision		
		Do not reject H_0	**Reject H_0**	**Total**
Actual Setting	**True H_0**	U	V	m_0
	False H_0	T	S	$m - m_0$
	Total	$m - R$	R	m

The Multiple Hypothesis Problem

Benjamini and Hochberg [1995] defined FDR as

$$FDR = E\left[\frac{V}{max(R,1)}\right] = E\left[\frac{V}{R}\right] \text{ for } R \geq 1 \tag{1}$$

The Multiple Hypothesis Problem

The FDR has also been investigated on a Bayesian perspective, as posterior probability distribution holding certain assumptions on the said framework.

Consider the problem of testing m hypotheses. Let H_i be equal to 0 if the null hypothesis is true, 1 if it is false.

Note that $H_i \sim Bernoulli(1 - \pi_0)$, where π_0 is our prior knowledge of the proportion of the null hypotheses being tested that are actually true (i.e. $\pi_0 = m_0/m$).

Then π_0 also happens to be the probability a null hypothesis being true, which we write as $P(H_i = 0)$.

The Multiple Hypothesis Problem

Define $\pi_1 = 1 - \pi_0$, the probability of null hypothesis being false, $P(H_i = 1) = 1 - P(H_i = 0)$.

Furthermore, for the m corresponding test statistics $x_1, x_2, ., x_m$, let f_0 be the density of x_i, if the null hypothesis is true; if it is false, let f_1 be its density.

Then the distribution of X is given by

$$f(x) = \pi_0 f_0(x) + \pi_1 f_1(x) \tag{2}$$

The Multiple Hypothesis Problem

With the assumptions discussed above, Storey [2003] presented the form of Bayesian FDR (FDR) as follows:

$$FDR(\mathcal{R}) = P(H = 0 | x \in \mathcal{R}) \tag{3}$$

where $\mathcal{R}$ is the set of "rejected" null hypotheses.

$FDR(\mathcal{R})$ represents the resulting false discovery rate of choosing to declare the tests in $\mathcal{R}$ as significant.

The Multiple Hypothesis Problem

Now let F_0 be the cumulative distribution function (CDF) corresponding to f_0 , and F be the CDF corresponding to f.

Storey [2003] showed that by applying Bayes' rule, the FDR takes the form

$$FDR(\mathcal{R}) = \frac{\pi_0 F_0(\mathcal{R})}{F(\mathcal{R})} \tag{4}$$

The Multiple Hypothesis Problem

Efron [2005] further explored this Bayesian interpretation of the FDR. While FDR comprises the whole set of hypotheses, Efron introduced the local fdr, which is the probability that a hypothesis is true given the data, which is given by

$$fdr(x) = \frac{\pi_0 f_0(x)}{f(x)} \tag{5}$$

The Multiple Hypothesis Problem

Methods for estimating these figures have already been explored for the continuous case. Efron and Tibshirani [2002] let f0 be known and assumed to follow the standard normal $N(0,1)$ distribution. The CDF $F(\mathcal{R})$ is then estimated by

$$\widehat{F}(R) = \frac{\#\{z_i \in \mathcal{R}\}}{m} \tag{6}$$

The Multiple Hypothesis Problem

As for the local fdr, a close analog is the following estimator:

$$\widehat{fdr}(x) = \frac{\pi_0 f_0(x)}{\widehat{f}(x)} \tag{7}$$

Using Discrete Kernels

Kernel Density Estimation is a nonparametric approach, which attempts to improve upon standard histogram methods by reweighting the data with the use of "kernel functions." A rough analog of the histogram method may be presented as follows:

$$\widehat{f}(x) = \frac{1}{n} \sum_{j} I(x_j = x) \tag{8}$$

Using Discrete Kernels

The kernel function will have to be replaced with one that is more suited for a discrete support. We refer to two such kernels, the first being the Dirac kernel mentioned in Li and Racine[2007]:

$$k_j(x; h) = \begin{cases} 1 - h & x = x_j \\ \frac{h}{c-1} & x \neq x_j \end{cases} \tag{9}$$

as well as the Aitken kernel defined in Li et al [2008]:

$$k_j(x; h) = \begin{cases} 1 & x = x_j \\ h & x \neq x_j \end{cases} \tag{10}$$

Using Discrete Kernels

Algorithm 1 Estimating the Empirical Bayes FDR

Initialize: $\hat{\tau}_j = 1 - \pi_0$
while $err \leq \epsilon$ **do**
$\widehat{f_1}(x) \leftarrow \sum_j \tau_j K(x_j; h) / \sum_i \tau_i$
$\widehat{f}(x) \leftarrow \pi_0 f_0(x) + (1 - \pi_0)\widehat{f_1}(x)$
$\hat{\tau}_j \leftarrow 1 - \pi_0 f_0(x) / \widehat{f}(x)$
$err \leftarrow \max_j \frac{|\tau_j^{(l)} - \tau_j^{(l-1)}|}{\tau_j^{(l-1)}}$
end while

Using Discrete Kernels

Algorithm 2 Cross-validation on the Kernel Smoothing Parameter

Initialize: randomly divide set of values x into V non-overlapping folds: $Y_1, Y_2, ..., Y_V$. Define X_k as the collection of all $Y_{i \neq k}$, that is the data-set taking out the kth fold.
for fold $i \in \{1, 2, ..., V\}$ **do**
Fit $\hat{f}_1$ on X_i (the training set) using Algorithm 1 with given h
Using the fitted $\hat{f}_1$, estimate $\hat{f}_1(y)$ for every y from Y_i (the testing set)
Define the log-likelihood of the subset y_i as $L(y_i; h) = \sum_j \ln \hat{f}_1(y_j)$
Then the V-fold cross-validation log-likelihood is defined as $L_{CV}(h) = \frac{1}{V} \sum_j L(y_j; h)$
end for
We choose $h = \arg\max_h L_{CV}(h)$

Using Discrete Kernels

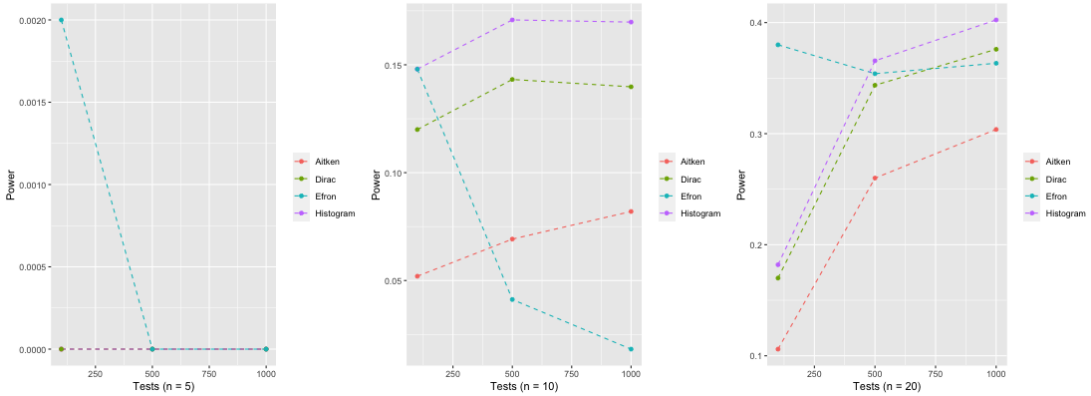


Figure: Simulated power, assuming true $\pi_0 = 0.95$, in the mixed case

Using Discrete Kernels

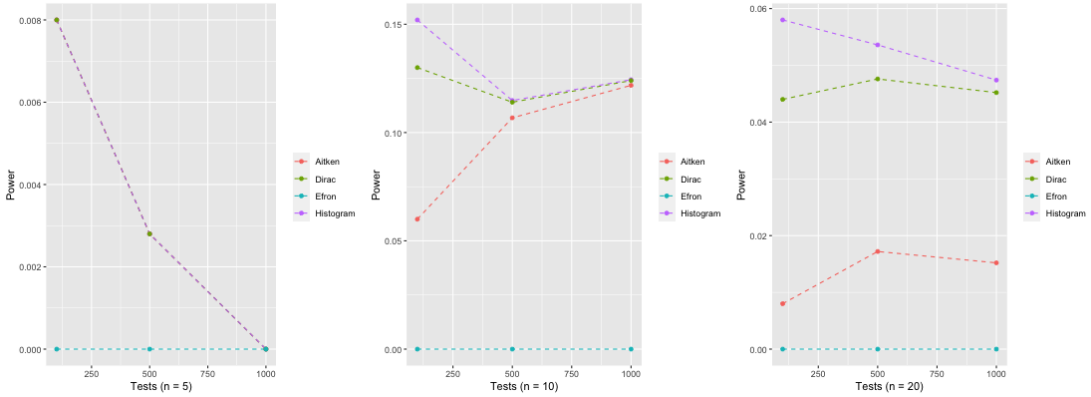


Figure: Simulated power, assuming true $\pi_0 = 0.95$, in the overlapping case

Using Discrete Kernels

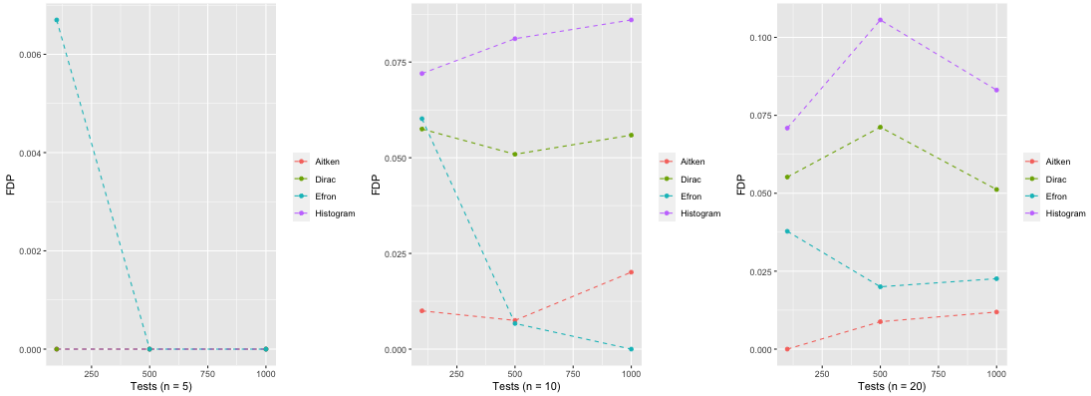


Figure: Simulated FDP, assuming true $\pi_0 = 0.95$, in the mixed case

Using Discrete Kernels

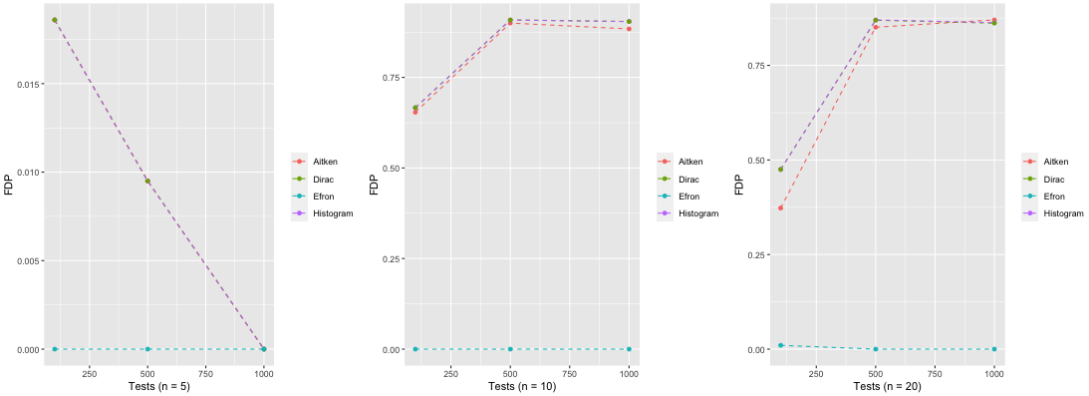


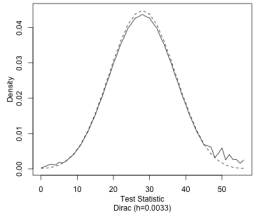
Figure: Simulated FDP, assuming true $\pi_0 = 0.95$, in the overlapping case

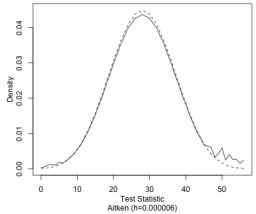
Application to Hedenfalk [2001]

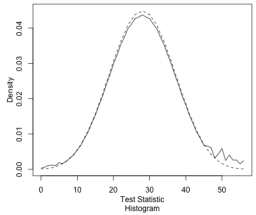
We apply the proposed procedure on data reported by Hedenfalk et al. [2001], concerning patients diagnosed with two different kinds of breast cancer.

The data used in the present paper concerns seven BRCA1 and eight BRCA2 patients (corresponding to gene mutations characterizing the disease), whose expression ratios were measured across $m = 3226$ genes.

Application to Hedenfalk [2001]







Application to Hedenfalk [2001]

A total of 51 genes were found to be significant at $FDR_i < 0.10$ if estimation using the Dirac kernel is applied.

We observe that Efron's procedure tends to overestimate not only the proportion of true nulls but also the corresponding local false discovery rates.

Cutoff, q	Genes with fdr <q	
	Dirac	Efron
0.050	0.00	0.00
0.075	26.00	0.00
0.100	51.00	0.00

Application to Hedenfalk [2001]

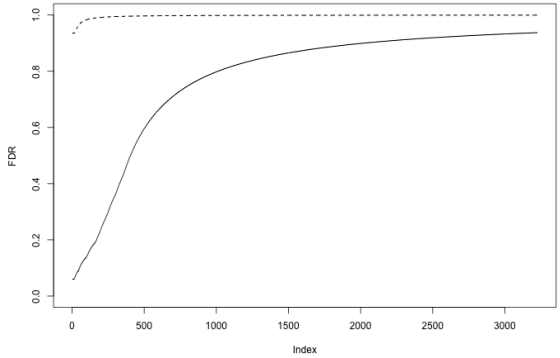


Figure: Estimated Bayesian FDR (ordered) for Dirac (solid) versus Efron (dashed)

Selected References

1. Efron, B (2004). "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis." Journal of the American Statistical Association. 2004, 99:96-104.
2. Chen, Xiongzhi and Doerge, R.W. (2015). "A weighted FDR procedure under discrete and heterogeneous null distributions." [URL](https://arxiv.org/abs/1502.00973)
3. Guedj, Mickael; Robin, Stephane; Celisse, Alain; and Nuel, Gregory (2009). "Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation." BMC Bioinformatics 2009, 10:84. DOI:10.1186/1471-2105-10-84
4. Li, Q.; and Racine, J. (2007). "Nonparametric econometrics: Theory and practice." Princeton University Press.
5. Benjamini, Y; and Hochberg, Y (1995): "Controlling the false discovery rate: a practical and powerfull approach to multiple testing." JRSSB 1995, 57:289-300.
6. Hedenfalk, I; Duggan, D; Chen, YD; Radmacher, M; Bittner, M; Simon, R; Meltzer, P; Gusterson, B; Esteller, M; Kallioniemi, OP; Wilfond, B; Borg, A; and Trent, J (2001). "Gene expression profiles in hereditary breast cancer." New England Journal of Medicine, 344:539-548.