arXiv:1106.3885v1 [stat.ME] 20 Jun 2011

A nonparametric empirical Bayes framework for large-scale multiple testing

Ryan Martin
Department of Mathematical Sciences
Indiana University–Purdue University Indianapolis
rgmartin@math.iupui.edu

Surya T. Tokdar
Department of Statistical Science
Duke University
st118@stat.duke.edu

May 2, 2011

Abstract

We propose a flexible and identifiable version of the two-groups model, motivated by hierarchical Bayes considerations, that features an empirical null and a semiparametric mixture model for the non-null cases. We use a computationally efficient predictive recursion marginal likelihood procedure to estimate the model parameters, even the nonparametric mixing distribution. This leads to a nonparametric empirical Bayes testing procedure, which we call PRtest, based on thresholding the estimated local false discovery rates. Simulations and real-data examples demonstrate that, compared to existing approaches, PRtest's careful handling of the non-null density can give a much better fit in the tails of the mixture distribution which, in turn, can lead to more realistic conclusions.

Keywords and phrases: Dirichlet process; marginal likelihood; mixture model; predictive recursion; two-groups model.

1 Introduction

Large-scale multiple testing problems arise in many applied fields such as genomics (Dudoit and van der Laan 2008; Schäfer and Strimmer 2005), proteomics (Ghosh 2009), astrophysics (Liang et al. 2004; Miller et al. 2001), and image analysis (Lindquist 2008; Schwartzman, Dougherty, and Taylor 2008), to name a few. An abstract representation of the problem is testing a set of hypotheses

$$H_{0i} : \text{the } i^{\text{th}} \text{ case manifests a "null" behavior}, \quad i = 1, \ldots, n$$

based on summary test statistics, or z-scores, $Z_1, \ldots, Z_n$. The null behavior of a single z-score Z_i can be described by the $\mathsf{N}(0,1)$ distribution when Z_i is defined as the Gaussian transform of a test statistic derived for the i^{th} case, such as the two sample t-statistic comparing treatment to control. Although this characterization leads to a simple rejection rule for the i^{th} case in isolation, it is found insufficient when all n tests in are to be performed simultaneously, particularly when n is very large. In fact, one of the major developments of modern statistics has been the philosophical shift from treating the z-scores as mutually independent to treating them as exchangeable (Efron and Tibshirani 2002). Consequently, recent work on large-scale simultaneous testing has focused on Bayesian models and, in particular, empirical Bayes methods that allow for information sharing between cases, even though separate decisions will be made for each case.

An elegant formalization of the large-scale simultaneous testing problem is the *two-groups model* (Efron 2004, 2007, 2008) which assumes $Z_1, \ldots, Z_n$ arise from a mixture density

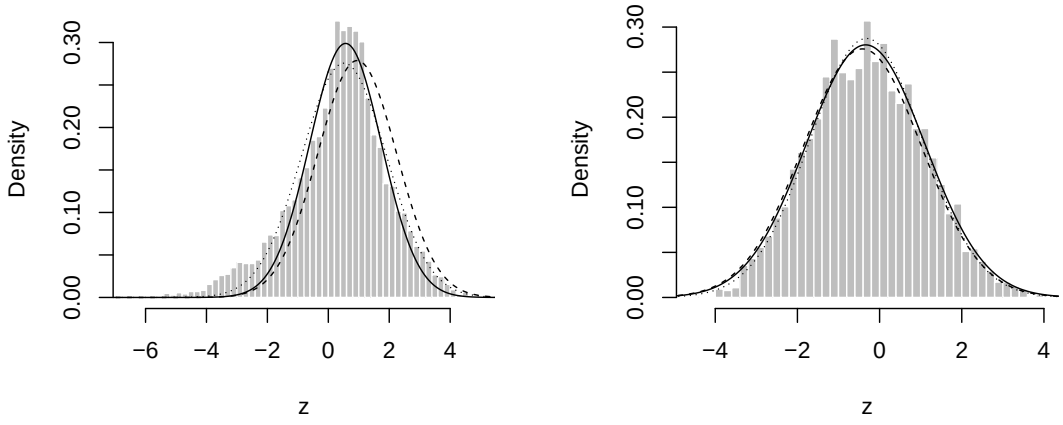
$$m(z) = \pi m_0(z) + (1-\pi) m_1(z), \tag{1}$$

with m_0 and m_1, respectively, describing the null and non-null distributions of the z-scores. Efron (2004, 2008) argues that, within the exchangeable setting, the case-specific theoretical null distribution $\mathsf{N}(0,1)$ may not be an adequate choice for m_0. A more appropriate choice is the so-called empirical null distribution $\mathsf{N}(\mu, \sigma^2)$, where μ and σ are to be estimated from data.

Following Efron's original treatment, various new methods have been proposed for fitting and drawing inference from the two groups model of z-scores (Jin and Cai 2007; Muralidharan 2010). These methods, together with related methodology based on p-values or t-scores (e.g., Benjamini and Hochberg 1995; Storey 2003), have been widely used in biological studies with high-throughput data, in particular to identify genes responsible for a phenotypical behavior based on microarray analysis. The single-summary-per-case approach of these methods offers substantial computational advantage over other approaches to analyze such data, such as those based on high-dimensional classification techniques (Golub et al. 1999; Lee et al. 2003).

However, currently available methods for fitting (1) do not take full advantage of the two-groups formulation. Motivated by applications to microarray studies, where typically a very small fraction of genes are linked with the phenotype, existing two-groups methods take a conservative approach of encouraging estimates of π close to 1. While this is reasonable for many applications, there are scientific studies where such a conservative approach fails to detect any or a majority of the interesting cases. Figure 1 reports two such microarray studies, a leukemia study by Golub et al. (1999) and a breast cancer study by Hedenfalk et al. (2001); more details are given in Section 6. As shown in the figure, existing methods each produce estimates of the null component πm_0 that cover one or both tails of the z-score histogram, leaving little to be explained by the non-null component $(1-\pi)m_1(z)$. Consequently, zero discoveries of interesting genes are made in one or both tails; see Table 2 in Section 6. High-dimensional classification-based analyses (Golub et al. 1999; Hedenfalk et al. 2001; Lee et al. 2003), on the other hand, identify a number of interesting genes on either tail for each of the two studies.

In this paper we consider a new likelihood-based analysis of the two-groups model, with a regularization on μ, σ, π and a semiparametric specification of the non-null density m_1. We employ a mixture representation of m_1 that gives it heavier tails than m_0 to



(a) Leukemia z-scores

(b) Breast cancer z-scores

Figure 1: Density histogram of z-scores from leukemia microarray data (Golub et al. 1999) and breast cancer data (Hedenfalk et al. 2001) along with estimates of $\pi m_0(z)$ based on the methods of Efron (2004) (—), Jin and Cai (2007)(−−) and Muralidharan (2010) $(\cdots)$.

reflect the belief that z-scores from the non-null cases are likely to be larger in magnitude than those from the null cases. The null weight π is given a beta prior with a center close to one but with a relatively long left tail. Additionally, we use a prior on (μ, σ) to reflect the belief that this vector is likely to be close to $(0, 1)$.

Compared to the existing methods based on z-scores, our proposal allows a wider range of estimates of π. For scientific studies where the existing methods discover a fair number of interesting cases, our method makes similar discoveries. But for other studies where existing methods fail, such as the two studies mentioned earlier, our method makes discoveries that are comparable to those found via high-dimensional classification methods. A similar adaptability property manifests in a simulation study where z-scores are generated according to (1)) with π ranging between 0.75 to 0.99; see Table 1 in Section 5.

Despite a nonparametric specification of m_1 and a likelihood-based analysis, our treatment of the two-groups model retains the computational efficiency that is hallmark of methods based on z-scores. This has been possible due to recent developments on a stochastic algorithm due to Newton (2002) called *predictive recursion*, for estimation of mixing densities with respect to any arbitrary dominating measure; see also Newton, Quintana, and Zhang (1998) and Newton and Zhang (1999). Theoretical properties of this algorithm are addressed in Ghosh and Tokdar (2006), Martin and Ghosh (2008), Tokdar, Martin, and Ghosh (2009), and Martin and Tokdar (2009). Martin and Tokdar (2011) show how this algorithm can be used in a hierarchical mixture model to construct a likelihood function over non-mixing model parameters, marginalized over the mixing density. This marginal likelihood is shown to have strong connections to the marginal

likelihood under a Bayesian Dirichlet process mixture model. We adopt this marginal likelihood calculation to the two groups model, with μ, σ, π and a scaling parameter in the specification of m_1 serving as the non-mixing parameters.

For the multiple testing problem, we adopt the strategy of mimicking the Bayes oracle rule by thresholding a plug-in estimate of the local false discovery rate, similar to Efron (2004, 2008), Jin and Cai (2007), and Muralidharan (2010). Simulations presented in Section 5 show that the proposed method, called PRtest, is more adaptive to asymmetry in the non-null density m_1 and the degree of sparsity characterized by π compared to these existing methods. Performance of PRtest in the two real-data applications described in Figure 1 is discussed in Section 6. There we find that the PR-based estimation produces a better fit in the tails of the distribution than that seen in Figure 1 and, consequently, we are able to identify a number of interesting genes in each example. The identified genes are, in fact, consistent with those identified by high-dimensional classification-based techniques.

2 Model specification

We take $m_0(z) = \mathsf{N}(z \mid \mu, \sigma^2)$, the normal density with unknown mean and variance μ and σ^2. The non-null density m_1 is taken to be a semiparametric mixture of the form

$$m_1(z) = \int_{\mathscr{U}} \mathsf{N}(z \mid \mu + \tau\sigma u, \sigma^2)\varphi(u)\, du, \tag{2}$$

with φ a density with respect to the Lebesgue measure on $\mathscr{U} = [-1, 1]$ and $\tau \geq 1$ a scaling factor. An important consequence of the requirement that φ be a density is given in the following theorem; see Appendix A for a proof.

Theorem 1. *For m_0 and m_1 as described above, the parameters $(\mu, \sigma, \pi, \tau, \varphi)$ in our version of the two-groups model are identifiable.*

This result is useful because, in general, identifiability is not guaranteed for a two groups model (1) with an empirical null that involves unknown parameters. For our specification, the key to identifiability is the model feature that m_1, by virtue of averaging over locations shifts of m_0, has heavier tails than m_0 itself. This feature is scientifically relevant as it embeds the belief that z-scores in the tails of the histogram are more likely to come from the non-null component than the null. Efron's method incorporates a similar belief through the *zero assumption* (Efron 2008) that most z-scores near zero are from the null component. However, such a zero assumption can be too strong to allow learning from data and can lead to an estimate of πm_0 that has heavier tails than any reasonable histogram-smoothing estimate of m, as reported by Strimmer (2008) and illustrated in Figure 1. In comparison, separating m_0 and m_1 by their tails seems more practical; see Section 6.

3 Mixture models and predictive recursion

It is more convenient to write our specification of m as the mixture model

$$m(z) = \int_{\mathscr{U}} p(z \mid \theta, u)\, F(du) \tag{3}$$

with parameters $\theta = (\mu, \sigma, \tau)$, kernel $p(z \mid \theta, u) = \mathsf{N}(z \mid \mu + \tau\sigma u, \sigma^2)$, and mixing probability measure F on $\mathscr{U}$ that assigns a positive mass π at $0 \in \mathscr{U}$ and distributes the remaining mass on $\mathscr{U}$ according to a Lebesgue density φ. The collection of all such F is the set $\mathbb{F} = \mathbb{F}(\mathscr{U}, \nu)$ of probability measures that are absolutely continuous with respect to the measure ν, defined to be the sum of the Lebesgue measure on $\mathscr{U}$ and a point mass at 0. The ν-density of such an F will be denoted by $\pi\langle 0\rangle + (1-\pi)\varphi$.

Inference on (θ, F) can be performed in a Bayesian setting with a prior distribution on (θ, F). A popular choice of prior distribution for the non-parametric probability measure F is the Dirichlet process prior (Ferguson 1973; Lo 1984). However, there are two practical difficulties in employing this inference framework for our model. First, the Dirichlet process prior entertains only discrete probability measures, thus violating the important absolute continuity property of F with respect to ν. Second, despite recent advances in computing, fitting a Dirichlet process mixture model does not scale well with the number of observations n. For microarray studies, n ranges from thousands to tens of thousands, whereas for more recent single nucleotide polymorphism studies, n equals a few hundreds of thousands. For such massive data sets, fitting a Dirichlet process mixture model can be fairly time consuming, nullifying some of the advantages of the two-groups framework.

As an alternative, we estimate (θ, F) via the predictive recursion (PR) methodology (Martin and Tokdar 2011; Newton 2002). Predictive recursion (Newton 2002) is a stochastic algorithm for estimating a mixing distribution through fast, recursive updates that have a strong connection with posterior updates for Dirichlet process mixture models. The algorithm accommodates user-specified absolute continuity constraints on the mixing distribution and enjoys attractive convergence properties under mild conditions with allowance for model misspecification (Ghosh and Tokdar 2006; Martin and Ghosh 2008; Martin and Tokdar 2009; Tokdar et al. 2009). However, Newton's original proposal can estimate the mixing distribution only when the kernel being mixed is known exactly, i.e., for (3), an estimate of F is available only when θ is known. To resolve this difficulty, Martin and Tokdar (2011) introduce a "marginal likelihood" function over non-mixing parameters, such as θ, based on the output of the predictive recursion. For any θ, the following calculations are performed.

PR Algorithm. Start with an initial estimate F_0 with ν-density $\pi_0\langle 0\rangle + \bar{\pi}\varphi_0$ and a sequence of weights $w_1, \ldots, w_n \in (0,1)$. For $i = 1, \ldots, n$ compute

$$m_{i-1,\theta}(Z_i) = \int p(Z_i \mid \theta, u)\, F_{i-1}(du),$$

$$F_i(du) = (1-w_i)F_{i-1}(du) + w_i \frac{p(Z_i \mid \theta, u)F_{i-1}(du)}{m_{i-1,\theta}(Z_i)}. \tag{4}$$

Produce $F_n = \pi_n\langle 0\rangle + \bar{\pi}_n\varphi_n$ as an estimate of F and $L_n(\theta) = \prod_{i=1}^n m_{i-1,\theta}(Z_i)$ as the marginal likelihood of θ.

Martin and Tokdar (2011) point out several justifications for labeling $L_n(\theta)$ as a likelihood function of θ. For $n = 1$, $L_1(\theta)$ equals the marginal likelihood function of θ, integrating out F under the Bayesian specification $F \sim \mathsf{DP}(\alpha, F_0)$, the Dirichlet process distribution with precision $\alpha = (1-w_1)/w_1$ and base measure F_0. For $n > 1$, this correspondence is not exact, but $L_n(\theta)$ can be viewed as a filtering approximation of the

corresponding Dirichlet process marginal likelihood function. Additionally, $L_n(\theta)$ features an asymptotic concentration property commonly enjoyed by likelihood functions for independent and identically distributed data models (Wald 1949). Specifically, for large n, with $Z_1, \cdots, Z_n$ independently drawn from a common density $m^\star$, $\log L_n(\theta) \approx -nK^\star(\theta)$, where $K^\star(\theta)$ equals the minimum Kullback–Leibler divergence between $m^\star$ and densities m of the form (3) with F ranging over the set $\mathbb{F}$ and all its weak limits points.

4 Regularized predictive recursion inference and PRtest

We employ a regularized version of the predictive recursion methodology to estimate θ and F for our two groups model. The regularization is motivated by a hierarchical Bayes formulation of (3) with $F \sim \mathsf{DP}(\alpha, F_0)$ where hyper-prior distributions are specified on the model parameters μ, σ, τ and F_0. We take the ν-density of F_0 to be $\pi_0\langle 0\rangle + (1-\pi_0)\varphi_0$ with a fixed choice of $\varphi_0(u) \propto u^2$, $u \in \mathscr{U}$. Among the remaining parameters, $\sigma \in (0, \infty)$, $\tau \in (1, \infty)$ and $\pi_0 \in (0, \infty)$ are taken to be independent with $\log \sigma \sim \mathsf{N}(0, 0.25^2)$, $\log(\tau - 1) \sim \mathsf{N}(0, 1)$ and $\pi_0 \sim \mathsf{Beta}(22.7, 1)$. Given σ and the other parameters, μ is assigned the conditional prior distribution $\mathsf{N}(0, \sigma^2/400)$.

In our experience, σ in the range $[0.5, 2.0]$ is typical, and the log-normal prior puts nearly all of its mass there. Other priors for σ may also be considered, such as a conjugate scaled inverse-chi distribution. The restriction $\tau > 1$ ensures that the non-null density m_1 is considerably wider than m_0, and the normal prior for $\log(\tau - 1)$ supports a large set of values in this range. The 22.7 in the beta prior for π_0, also used by Bogdan, Ghosh, and Tokdar (2008), assigns about 90% of its mass to the interval $[0.9, 1]$, reflecting the belief that the null proportion π is likely to be large. Finally, the prior for μ is scaled to the choice of σ and highly concentrated around the origin, reflecting the belief that the z-scores should have mean close to zero. Finer tuning of this default prior for specific problems is straightforward.

For a predictive recursion analog of this hierarchical Bayesian model, we interpret the predictive recursion likelihood as a function of both $\theta = (\mu, \sigma, \tau)$ and π_0. Writing this likelihood as $L_n(\mu, \sigma, \tau, \pi_0)$ and letting $g(\mu, \sigma, \tau, \pi_0)$ denote the joint prior density function on these parameters, a regularized version of the predictive recursion marginal log-likelihood function can be written as

$$\tilde{\ell}_n(\mu, \sigma, \tau, \pi_0) = \log L_n(\mu, \sigma, \tau, \pi_0) + \log g(\mu, \sigma, \tau, \pi_0). \tag{5}$$

Estimates of these parameters are obtained by maximizing $\tilde{\ell}_n = \tilde{\ell}_n(\mu, \sigma, \tau, \pi_0)$. Once these estimates are obtained, predictive recursion is run one last time with the estimated values of these parameters to produce an estimate of F, i.e., of π and of φ in (1) and (2), respectively. In our implementations, maximization of ℓ_n is done by the gradient-based optimization method known as the Broyden-Fletcher-Goldfarb-Shanno, or BFGS. In Appendix B we provide a variation on the PR algorithm that produces the gradient of $\log L_n$ as a by-product.

The predictive recursion methodology depends on two additional factors, namely, the choice of weights $w_1, \ldots, w_n$ and the order in which the z-scores are processed by the algorithm. Martin and Tokdar (2009) provide an upper bound on the rate of convergence for PR estimates of the mixture m when the weights are of the form $w_i = (1 + i)^{-\gamma}$,

$\gamma \in (2/3, 1]$. Our choice $w_i = (i+1)^{-0.67}$ is close to the limit $\gamma = 2/3$ where the upper bound is optimal. The recursive nature of the algorithm induces dependence on the order in which the Z_i values are visited. We reduce this dependence by replacing $\tilde{\ell}_n$ with its average over a number of random permutations of the data sequence. Averaging over permutations increases the overall computation time, but adds stability to parameter estimation (Tokdar et al. 2009). In our experience, averaging over 10 random permutations is sufficient to stabilize the estimates of θ, and the additional computation time required is negligible. To reduce variability due to random permutation, we keep the set of permutations fixed over the process of maximizing $\tilde{\ell}_n$.

For multiple testing, we consider the local false discovery rate (Efron 2004), given by

$$\mathrm{fdr}(z) = \pi m_0(z)/m(z),$$

which represents the posterior probability that a case with z-score $Z = z$ is null. Sun and Cai (2007) argue that the local false discovery rate, rather than the traditional p-values, is the fundamental quantity for multiple testing. Once estimation PR of $(\mu, \sigma, \tau, \pi, \varphi)$ is completed, a plug-in estimate $\widehat{\mathrm{fdr}}$ of fdr is readily available, and PRtest is implemented by thresholding $\widehat{\mathrm{fdr}}$. In our examples we use a threshold of $r = 0.1$ and declare case i as non-null if $\widehat{\mathrm{fdr}}(Z_i) < r$. This is equivalent to assuming that a Type I error is nine times as costly as a Type II error. This choice, used by Sun and Cai (2007, Sec. 5), is somewhat subjective, but sits between the choice $r = 0.2$ of Efron (2008) and Strimmer (2008) and the choice $r = 0.05$ of Jin and Cai (2007) and others.

5 Simulations

Here we investigate the performance of PRtest in several large-scale simulations where we can compare the results with the benchmark Bayes oracle test. The results will also be compared to those obtained from the Fourier-based method of Jin and Cai (2007) and the mixfdr method of Muralidharan (2010).

For $Z_1, \ldots, Z_n$, we assume independence and take the null density as $m_0(z) = \mathsf{N}(z \mid \mu, \sigma^2)$. Here we fix $n = 1000$, $\mu = 0$, and $\sigma = 1$. Four choices of m_1 are considered:

C1: $m_1(z) = \mathsf{N}(z \mid 0, \sigma^2 + \omega^2)$. Taking $\omega^2 = 13 \approx 2\sigma^2 \log n$ ensures the non-null z-scores are "detectable" (Donoho and Johnstone 1994). But, in our experience, the range of z-scores one finds in real data analysis is consistent with smaller signals, so we take $\omega^2 = 4$.

C2: $m_1(z) = 0.5 \int_2^4 \mathsf{N}(z \mid u, \sigma^2)\, du$. This choice, used by Muralidharan (2010) and Johnstone and Silverman (2004), exhibits asymmetry and has only slightly heavier tails than the null.

C3: $m_1(z) = 0.67\, \mathsf{N}(z \mid -3, 2) + 0.33\, \mathsf{N}(z \mid 3, 2)$. This one is asymmetric and a large portion of its mass is concentrated away from the origin.

C4: $m_1(z) = 0.25 \int_{[-4,-2]\cup[2,4]} \mathsf{N}(z \mid u, \sigma^2)\, du$. This is a symmetrized version of C2. A key feature of this choice is that the unobserved signals are bounded away from zero.

m_1	π	Jin-Cai	mixfdr	PRtest
C1	0.75	0.928 (0.019)	0.957 (0.009)	0.918 (0.017)
	0.80	0.929 (0.019)	0.965 (0.007)	0.930 (0.016)
	0.85	0.934 (0.018)	0.971 (0.006)	0.942 (0.014)
	0.90	0.945 (0.015)	0.980 (0.005)	0.960 (0.014)
	0.95	0.961 (0.011)	0.989 (0.003)	0.980 (0.010)
	0.99	0.978 (0.005)	0.995 (0.001)	0.995 (0.003)
C2	0.75	0.905 (0.015)	0.827 (0.016)	0.761 (0.017)
	0.80	0.874 (0.019)	0.860 (0.012)	0.804 (0.014)
	0.85	0.860 (0.023)	0.894 (0.009)	0.851 (0.013)
	0.90	0.869 (0.028)	0.927 (0.007)	0.896 (0.010)
	0.95	0.926 (0.017)	0.962 (0.005)	0.940 (0.009)
	0.99	0.984 (0.007)	0.991 (0.003)	0.980 (0.008)
C3	0.75	0.909 (0.013)	0.857 (0.017)	0.788 (0.016)
	0.80	0.886 (0.015)	0.881 (0.013)	0.828 (0.015)
	0.85	0.871 (0.021)	0.909 (0.011)	0.867 (0.014)
	0.90	0.886 (0.020)	0.937 (0.008)	0.903 (0.014)
	0.95	0.935 (0.012)	0.967 (0.005)	0.937 (0.013)
	0.99	0.980 (0.004)	0.991 (0.003)	0.982 (0.010)
C4	0.75	0.951 (0.007)	0.886 (0.035)	0.784 (0.066)
	0.80	0.934 (0.010)	0.897 (0.015)	0.814 (0.021)
	0.85	0.920 (0.015)	0.920 (0.010)	0.862 (0.018)
	0.90	0.908 (0.025)	0.948 (0.007)	0.901 (0.013)
	0.95	0.929 (0.017)	0.975 (0.005)	0.943 (0.012)
	0.99	0.980 (0.007)	0.995 (0.002)	0.992 (0.005)

Table 1: Mean (standard deviation) of the 500 estimates of π for the method of Jin and Cai (2007), the mixfdr method of Muralidharan (2010), and PRtest for the four m_1's described in Section 5.

For each of the four choices of m_1, we consider six choices of $\pi \in \{0.75, 0.8, 0.85, 0.9, 0.95, 0.99\}$, forming a total of 24 simulations settings. Each setting is replicated 500 times and the results are reported below. Each implementation of PR is run with weights $w_i = (1+i)^{-0.67}$ and the regularized likelihood $\tilde{\ell}_n$ is averaged over 10 permutations of the data sequence.

Table 1 summarizes the estimates of the null parameters π for each simulation setting. Estimates of (μ, σ) are similar (and accurate) across methods, models, and π values, so these results are omitted. From the table we find that the maximum PR marginal likelihood estimates are the most adaptive across the range of π values, specifically for choices C2–C4. Of particular interest is PRtest's strong performance in the practically realistic choice C3, having smooth bimodal non-null density. Also the average computation time for PRtest is roughly 3 seconds, which compares favorably with average computing times required by Jin-Cai (≈ 0.7 seconds) and mixfdr (≈ 0.5 seconds).

Next we compare the performance of the selected methods based on false non-discovery rate, false discovery rate, power, and Bayes risk. We limit this discussion to non-null choice C3 as it is arguably the most realistic and the results for the other models are

similar. Figure 2 plots these quantities as functions of π for the selected methods and the Bayes oracle. The general message is that PRtest is competitive with the other tests in all aspects across a range of sparsity levels. In particular, the four tests are similar in terms of false non-discovery rate, particularly for large π, but PRtest is better than mixfdr and Jin-Cai for relatively small π. Also, each of the four tests have relatively small false discovery rates, although the Jin-Cai method has a somewhat unexpected spike, which explains its higher power for large π values. The Bayes oracle test has the smallest Bayes risk uniformly over π, but the PRtest risk sits very close over the entire range of π. This observation suggests that our PR-based procedure may be asymptotically optimal in the sense that the ratio of the risk to the Bayes oracle risk approaches 1 as $(n, \pi) \to (\infty, 1)$ at a certain rate; see, e.g., Bogdan et al. (2011).

6 Real-data examples

We apply our PRtest method to microarray gene expression datasets from two scientific studies, a leukemia study by Golub et al. (1999) and a hereditary breast cancer study by Hedenfalk et al. (2001). In the leukemia study, z-sores were obtained for 7,129 human genes by comparing 27 patients with acute lymphoblastic leukemia against 11 patients with acute myeloid leukemia. In the breast cancer study z-scores for 3,228 human genes were obtained by comparing seven hereditary breast cancer patients with BRCA1 mutation against eight with BRCA2 mutation.

For these data sets, the methods by Efron (2004), Jin and Cai (2007) and Muralidharan (2010) fail to identify differentially expressed genes in either one or both tails of the z-score histogram, as shown in Table 2; see, also, Figure 1. However, classification-based methods identify interesting and biologically relevant genes in both tails (Golub et al. 1999; Hedenfalk et al. 2001; Lee et al. 2003).

Figure 3 gives a visual summary of our method's fit to these two data sets. Several discoveries are made with decision rule $\widehat{\text{fdr}} < 0.1$ in both tails of the z-score histogram for either data set. Also shown are the genes identified by Lee et al. (2003) and, additionally, for the leukemia data, Golub et al. (1999). The discoveries made by our method are clearly consistent with the discoveries made by these two classification based approaches. Table 2 compares our method with the three existing two groups methods through $\hat{\pi}$ and the number of genes identified in each tail of the z-scores histogram.

Software

Software to implement the proposed PRtest methodology can be found at the second author's website, `http://www.stat.duke.edu/~st118`.

Acknowledgments

The authors thank Professor J. K. Ghosh for helpful discussions.

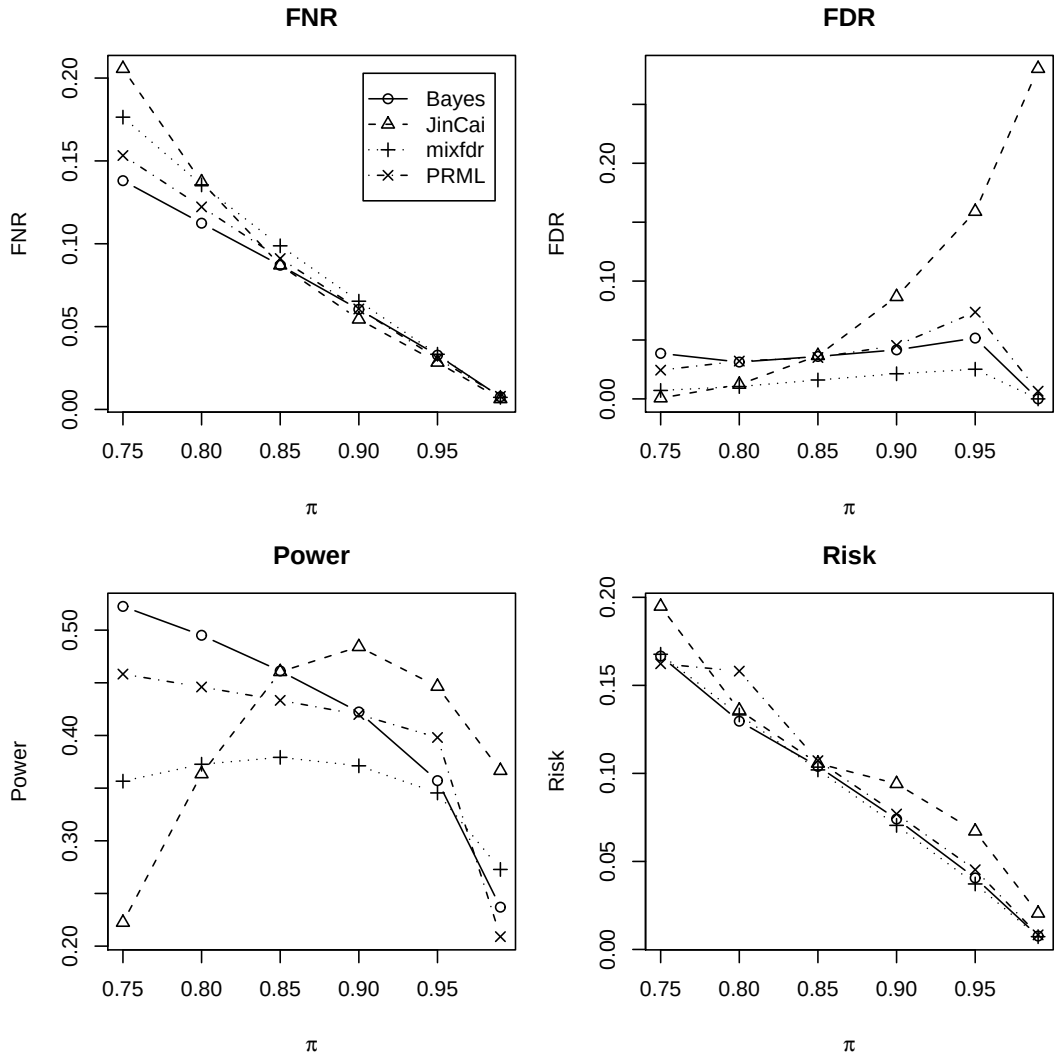
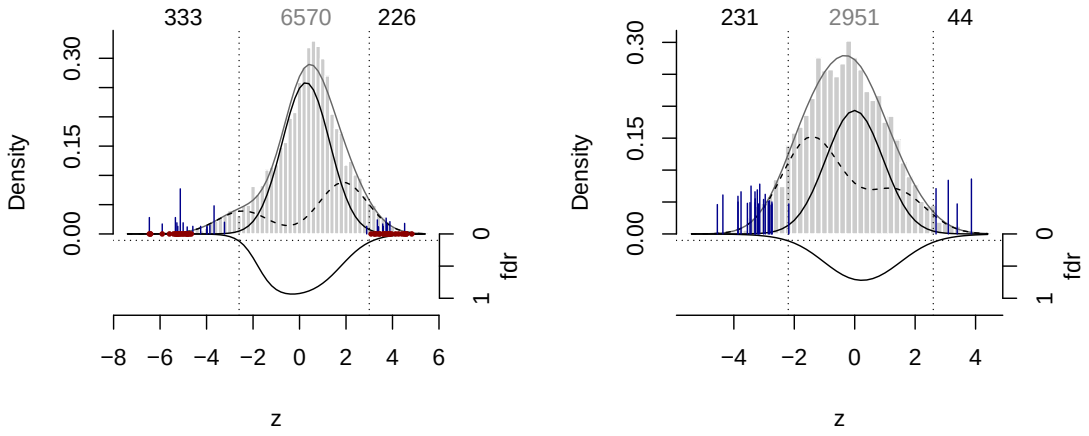


Figure 2: Plots of the false non-discovery rate (FNR, top left), false discovery rate (FDR, top right), power (bottom left), and Bayes risk (bottom right) against π for the selected testing procedures in the C3 simulation setting described in Section 5.

	Leukemia			*Breast Cancer*		
	$\hat{\pi}$	Left	Right	$\hat{\pi}$	Left	Right
Efron	0.88	276	0	1.00	0	0
Jin-Cai	0.91	291	0	1.00	0	0
mixfdr	0.96	71	0	0.99	0	0
PRtest	0.63	333	226	0.45	231	44

Table 2: A comparison of PRtest with the methods by Efron (2004), Jin and Cai (2007) and Muralidharan (2010) for leukemia and breast cancer gene expression data. Methods are compared based on the null proportion estimate ($\hat{\pi}$) and the number of significant genes identified in the left tail (Left) and in the right tail (Right) of the z-scores histogram in Figure 3.



(a) Leukemia

(b) Breast cancer

Figure 3: PRtest's fit to z-scores values from leukemia microarray data and breast cancer data. Overlaid on the z-score histogram are the estimates of πm_0 (solid, black), $(1-\pi)m_1$ (dashed, black) and $m = \pi m_0 + (1-\pi)m_1$ (solid, grey). The estimated fdr curve is shown on the negative scale, with the cut-off of 0.1 marked by the horizontal grey, dashed line. Genes identified by Lee et al. (2003) are marked with a blue bar at their z-score, with the height of the bar indicating the posterior probability of being included in the classification model. For the leukemia data, red dots are placed on z-scores of the genes identified in the original study by Golub et al. (1999).

A Proof of Theorem 1

Here we prove a more general version of Theorem 1 in the main text. Let $p(z)$ be a probability density function on $\mathbb{R}$, symmetric about zero. Furthermore, assume p is supersmooth in the sense of Fan (1991); see (9) below. In the main text, we took $p(z)$ to be a $\mathsf{N}(0,1)$ kernel but, e.g., a Student-t kernel with known degrees of freedom would also satisfy these conditions.

For the particular choice of p, define the mapping

$$M(\mu,\sigma,\tau,\pi,f)(z) = \pi\sigma^{-1}p\big((z-\mu)/\sigma\big) + (1-\pi)\int \sigma^{-1}p\big((z-\mu-\tau u)/\sigma\big)\varphi(u)\,du.$$

To prove that $(\mu,\sigma,\tau,\pi,\varphi)$ are identifiable, we need to show that M is a one-to-one function. Therefore, we start by assuming

$$M(\mu_1,\sigma_1,\tau_1,\pi_1,\varphi_1) = M(\mu_2,\sigma_2,\tau_2,\pi_2,\varphi_2).$$

Let $p^*(t)$ and $\varphi_k^*(t)$ denote the characteristic functions of $p(z)$ and $\varphi_k(u)$, respectively, for $k=1,2$. Then we must have

$$\begin{aligned}\exp\{it\mu_1\}p^*(\sigma_1 t)\big[\pi_1 + (1-\pi_1)\varphi_1^*(\sigma_1 t/\tau_1)\big] \\ = \exp\{it\mu_2\}p^*(\sigma_2 t)\big[\pi_2 + (1-\pi_2)\varphi_2^*(\sigma_2 t/\tau_2)\big]\end{aligned} \tag{6}$$

for every $t\in\mathbb{R}$, where i is the imaginary unit. Recall that the Riemann-Lebesgue lemma (e.g., Billingsley 1995, Theorem 26.1) says, for $k=1,2$,

$$\varphi_k^*(t)\to 0 \quad \text{as} \quad t\to\pm\infty. \tag{7}$$

Now, suppose $\sigma_1>\sigma_2$ and assume, without loss of generality, that $\mu_2>0$. Choose a sequence $\{t_s\}\subset\mathbb{R}$ such that $t_s\to\infty$ and $\exp\{it_s\mu_2\}\equiv 1$. Then, for large enough s, (7) would imply that $\pi_2+(1-\pi_2)\varphi_2^*(\sigma_2 t_s/\tau_2)\neq 0$. On rearranging the terms in (6) we get

$$\frac{p^*(\sigma_2 t_s)}{p^*(\sigma_1 t_s)} = \frac{\exp\{it_s\mu_1\}\big[\pi_1+(1-\pi_1)\varphi_1^*(\sigma_1 t_s/\tau_1)\big]}{\pi_2+(1-\pi_2)\varphi_2^*(\sigma_2 t_s/\tau_2)}. \tag{8}$$

We have assumed that p is supersmooth (Fan 1991), which means that

$$d_0|t|^{\beta_0}\exp\{-|t|^\beta/\gamma\} \le |p^*(t)| \le d_1|t|^{\beta_1}\exp\{-|t|^\beta/\gamma\}, \tag{9}$$

for all t and for some positive constants $d_0, d_1, \beta_0, \beta_1, \beta$, and γ. Under this assumption, the modulus of the left-hand side of (8) satisfies

$$\Big|\frac{p^*(\sigma_2 t_s)}{p^*(\sigma_1 t_s)}\Big| \ge \text{const}\times|t_s|^{\beta_1-\beta_0}\exp\{|t_s|^\beta(\sigma_1^\beta-\sigma_2^\beta)/\gamma\}.$$

Therefore, as $s\to\infty$, the left-hand side of (8) is unbounded while the right-hand side is bounded. This is a contradiction, so we need $\sigma_1\le\sigma_2$. But by symmetry, it follows that $\sigma_1=\sigma_2$. With this equality, relation (6) easily leads to the equalities $\mu_1=\mu_2$, $\tau_1=\tau_2$, $\pi_1=\pi_2$ and $\varphi_1=\varphi_2$.

B Gradient of the log PR marginal likelihood

This section provides a variation on the predictive recursion (PR) algorithm that yields the gradient of the log PR marginal likelihood function, based on the development in (Martin and Tokdar 2011). The model under consideration here is the following:

$$m(z) = \pi \mathsf{N}(z \mid \mu, \sigma^2) + \bar{\pi} \int \mathsf{N}(z \mid \mu + \tau\sigma u, \sigma^2) \varphi(u) \, du,$$

where φ is an unknown mixing density supported on $[-1, 1]$. The details of the PRtest method can be found in the main text. Here we focus only on computing the gradient of

$$\ell_n(\theta) = \sum_{i=1}^{n} \log m_{i-1,\theta}(Z_i),$$

where $m_{k,\theta}(z)$ is the PR estimate of the mixture density based on $Z_1, \ldots, Z_k$ and $\theta = (\mu, \sigma, \tau, \pi_0)$, slightly different than in the main text.

Define an unconstrained version of θ, i.e., $\eta = (\mu, \log\sigma, \log(\tau - 1), \operatorname{logit}\pi_0)$, where $\operatorname{logit} x = \log(\frac{x}{1-x})$. In what follows, ∇ will denote a gradient with respect to η, and if g is a function of a variable u, then $\nabla g(u)$ denotes the gradient with respect to η, pointwise in u. The following algorithm shows how to compute $\lambda_i = m_{i-1,\theta(\eta)}(Z_i)$ and $\nabla \log \lambda_i$ for $i = 1, \ldots, n$.

1. Start with user-specified π_0 and f_0, and set

$$\nabla\pi_0 = (0, 0, 0, \pi_0(1 - \pi_0)) \quad \text{and} \quad \nabla\varphi_0(u) \equiv (0, 0, 0, 0).$$

2. For $i = 1, \ldots, n$, repeat the following three steps:

   (a) For the normal kernel $p(z \mid \theta, u) = \mathsf{N}(z \mid \mu + \sigma\tau u, \sigma^2)$, set

$$G_0 = \mathsf{N}(Z_i \mid \mu, \sigma^2) \quad \text{and} \quad G_1(u) = \mathsf{N}(Z_i \mid \mu + \sigma\tau x, \sigma^2),$$

   and analytically evaluate the gradients ∇G_0 and $\nabla G_1(u)$:

$$\begin{aligned} \nabla G_0 &= (z_0/\sigma, z_0^2 - 1, 0, 0) \cdot G_0 \\ \nabla G_1(u) &= (z_1(u)/\sigma, z_1(u)\tau u/\sigma + z_1^2(u) - 1, z_1(u)u(\tau - 1), 0) \cdot G_1(u), \end{aligned}$$

   where $z_0 = (Z_i - \mu)/\sigma$ and $z_1(u) = (Z_i - \mu - \sigma\tau u)/\sigma$.

   (b) Compute

$$\begin{aligned} h_i &= \int G_1(u)\varphi_{i-1}(u) \, du \\ \lambda_i &= \pi_{i-1} G_0 + (1 - \pi_{i-1}) h_i \\ \nabla \log h_i &= \frac{1}{h_i} \int \big\{ G_1(u) \nabla\varphi_{i-1}(u) + \nabla G_1(u) \varphi_{i-1}(u) \big\} \, du \\ \nabla \log \lambda_i &= \frac{\nabla\pi_{i-1} G_0 + \pi_{i-1} \nabla G_0 + h_i \{ (1 - \pi_{i-1}) \nabla \log h_i - \nabla\pi_{i-1} \}}{u_i} \end{aligned}$$

(c) Update

$$\begin{aligned}
\pi_i &= A_0\pi_{i-1} \\
\nabla\pi_i &= A_0\nabla\pi_{i-1} + \nabla A_0\pi_{i-1} \\
\varphi_i(u) &= BA_1(u)\varphi_{i-1}(u) \\
\nabla\varphi_i(u) &= \{\nabla BA_1(u) + B\nabla A_1(u)\}\varphi_{i-1}(u) + BA_1(u)\nabla\varphi_{i-1}(u)
\end{aligned}$$

where

$$\begin{aligned}
A_0 &= 1 + w_i(G_0/\lambda_i - 1) \\
A_1(u) &= 1 + w_i(G_1(u)/\lambda_i - 1) \\
B &= (1 - \pi_{i-1})/(1 - A_0\pi_{i-1})
\end{aligned}$$

and

$$\begin{aligned}
\nabla A_0 &= w_i\{\nabla G_0 - G_0\nabla\log\lambda_i\}/\lambda_i \\
\nabla A_1(u) &= w_i\{\nabla G_1(u) - G_1(u)\nabla\log\lambda_i\}/\lambda_i \\
\nabla B &= \frac{(BA_0 - 1)\nabla\pi_{i-1} + B\nabla A_0\pi_{i-1}}{1 - A_0\pi_{i-1}}
\end{aligned}$$

3. Return the log-likelihood $\sum_{i=1}^n \log\lambda_i$ and its gradient $\sum_{i=1}^n \nabla\log\lambda_i$.

References

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

Billingsley, P. (1995), *Probability and measure*, New York: John Wiley & Sons Inc., 3rd ed.

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011), "The Bayes oracle and asymptotic optimality of multiple testing procedures under sparsity," *Ann. Statist.*, to appear.

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008), "A comparison of the Benjamini-Hochberg prodecure with some Bayesian rules for multiple testing," in *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, eds. Balakrishnan, N., Peña, E., and Silvapulle, M., Beachwood, OH: IMS, pp. 211–230.

Donoho, D. L. and Johnstone, I. M. (1994), "Minimax risk over l_p-balls for l_q-error," *Probab. Theory Related Fields*, 99, 277–303.

Dudoit, S. and van der Laan, M. J. (2008), *Multiple testing procedures with applications to genomics*, New York: Springer.

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

— (2007), "Correlation and large-scale simultaneous significance testing," *J. Amer. Statist. Assoc.*, 102, 93–103.

— (2008), "Microarrays, empirical Bayes and the two-groups model," *Statist. Sci.*, 23, 1–22.

Efron, B. and Tibshirani, R. (2002), "Empirical Bayes methods and False Discovery Rates for Microarrays," *Genet. Epidemiol.*, 23, 70–86.

Fan, J. (1991), "On the optimal rates of convergence for nonparametric deconvolution problems," *Ann. Statist.*, 19, 1257–1272.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, 1, 209–230.

Ghosh, D. (2009), "Assessing significance of peptide spectrum matches in proteomics: a multiple testing approach," *Statistics in Biosciences*, 1, 199–213.

Ghosh, J. K. and Tokdar, S. T. (2006), "Convergence and consistency of Newton's algorithm for estimating mixing distribution," in *Frontiers in statistics*, eds. Fan, J. and Koul, H., London: Imp. Coll. Press, pp. 429–443.

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.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O., B., W., Borg, A., Trent, J., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., and Sauter, G. (2001), "Gene-expression profiles in hereditary breast cancer," *N. Engl. J. Med.*, 344, 539–548.

Jin, J. and Cai, T. T. (2007), "Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons," *J. Amer. Statist. Assoc.*, 102, 495–506.

Johnstone, I. M. and Silverman, B. W. (2004), "Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences," *Ann. Statist.*, 32, 1594–1649.

Lee, K. E., Sha, N., Edward R. Dougherty and, M. V., and Mallick, B. K. (2003), "Gene selection: a Bayesian variable selection approach," *Bioinformatics*, 19, 90–97.

Liang, C.-L., Rice, J. A., de Pater, I., Alcock, C., Axelrod, T., Wang, A., and Marshall, S. (2004), "Statistical methods for detecting stellar occultations by Kuiper belt objects: the Taiwanese-American occultation survey," *Statist. Sci.*, 19, 265–274.

Lindquist, M. A. (2008), "The statistical analysis of fMRI data," *Statist. Sci.*, 23, 439–464.

Lo, A. Y. (1984), "On a class of Bayesian nonparametric estimates. I. Density estimates," *Ann. Statist.*, 12, 351–357.

Martin, R. and Ghosh, J. K. (2008), "Stochastic approximation and Newton's estimate of a mixing distribution," *Statist. Sci.*, 23, 365–382.

Martin, R. and Tokdar, S. T. (2009), "Asymptotic properties of predictive recursion: robustness and rate of convergence," *Electron. J. Stat.*, 3, 1455–1472.

— (2011), "Semiparametric inference in mixture models with predictive recursion marginal likelihood," *Biometrika*, to appear. Preprint `arXiv:1106.3352`.

Miller, C. J., Genovese, C., Nichol, R. C., Wasserman, L., Connolly, A., Reichart, D., and Hopkins, A. (2001), "Controlling false discovery rate in astrophysical data analysis," *Astron. J.*, 122, 3492–3505.

Muralidharan, O. (2010), "An empirical Bayes mixture method for effect size and false discovery rate estimation," *Ann. Appl. Statist.*, 4, 422–438.

Newton, M. A. (2002), "On a nonparametric recursive estimator of the mixing distribution," *Sankhyā Ser. A*, 64, 306–322.

Newton, M. A., Quintana, F. A., and Zhang, Y. (1998), "Nonparametric Bayes methods using predictive updating," in *Practical nonparametric and semiparametric Bayesian statistics*, eds. Dey, D., Müller, P., and Sinha, D., New York: Springer, vol. 133 of *Lecture Notes in Statist.*, pp. 45–61.

Newton, M. A. and Zhang, Y. (1999), "A recursive algorithm for nonparametric analysis with missing data," *Biometrika*, 86, 15–26.

Schäfer, J. and Strimmer, K. (2005), "An empirical Bayes approach to inferring large-scale gene association networks," *Bioinformatics*, 21, 754–765.

Schwartzman, A., Dougherty, R. F., and Taylor, J. E. (2008), "False discovery rate analysis of brain diffusion direction maps," *Ann. Appl. Stat.*, 2, 153–175.

Storey, J. D. (2003), "The positive false discovery rate: a Bayesian interpretation and the q-value," *Ann. Statist.*, 31, 2013–2035.

Strimmer, K. (2008), "A unified approach to false discovery rate estimation," *BMC Bioinformatics*, 9, 303.

Sun, W. and Cai, T. T. (2007), "Oracle and adaptive compound decision rules for false discovery rate control," *J. Amer. Statist. Assoc.*, 102, 901–912.

Tokdar, S. T., Martin, R., and Ghosh, J. K. (2009), "Consistency of a recursive estimate of mixing distributions," *Ann. Statist.*, 37, 2502–2522.

Wald, A. (1949), "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Statist.*, 20, 595–601.