Biostatistics (2012), **13**, 3, *pp.* 427–439
doi:10.1093/biostatistics/kxr039
Advance Access publication on November 14, 2011

A nonparametric empirical Bayes framework for large-scale multiple testing

RYAN MARTIN*

Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 S. Morgan Street, Chicago, IL 60607, USA
rgmartin@math.uic.edu

SURYA T. TOKDAR

Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708, USA

SUMMARY

We propose a flexible and identifiable version of the 2-groups model, motivated by hierarchical Bayes considerations, that features an empirical null and a semiparametric mixture model for the nonnull cases. We use a computationally efficient predictive recursion (PR) 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 nonnull density can give a much better fit in the tails of the mixture distribution which, in turn, can lead to more realistic conclusions.

Keywords: 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 *and others*, 2004; Miller *and others*, 2001), and image analysis (Schwartzman *and others*, 2008; Lindquist, 2008) to name a few. An abstract representation of the problem is testing a set of hypotheses

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

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 N(0, 1) distribution when Z_i is defined as the Gaussian transform of a test statistic derived for the ith case, such as the 2 sample t-statistic comparing treatment to control. Although this characterization leads to a simple rejection rule for the ith case in isolation, it is found insufficient when all n tests in are to be performed, 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 whom correspondence should be addressed.

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 "2-groups model" (Efron, 2004, 2007, 2008) which assumes $Z_1, \ldots, Z_n$ arise from a mixture density

$$f(z) = \pi f_0(z) + (1 - \pi) f_1(z), \tag{1.1}$$

with f_0 and f_1, respectively, describing the null and nonnull distributions of the z-scores. Efron (2004, 2008) argues that, for a variety of reasons, the case-specific theoretical null distribution $\mathrm{N}(0, 1)$ may not be an adequate choice for f_0, and a more appropriate choice is the so-called empirical null distribution $\mathrm{N}(\mu, \sigma^2)$, where μ and σ are to be estimated from data.

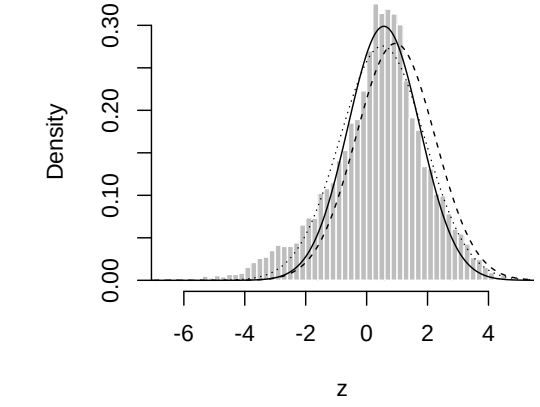
Following Efron's original treatment, various new methods have been proposed for fitting and drawing inference from the 2-groups model of z-scores (Jin and Cai, 2007; Muralidharan, 2010). These methods, together with related methodology based on p values or t-scores (e.g., Benjamini and Hochberg, 1995; 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, such as those based on high-dimensional classification (Golub *and others*, 1999; Lee *and others*, 2003).

However, currently available methods for fitting (1.1) do not take full advantage of the 2-groups formulation. Motivated by applications to microarray studies, where typically a very small fraction of genes are linked with the phenotype, existing 2-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 2 such microarray studies, a leukemia study by Golub *and others* (1999) and a breast cancer study by Hedenfalk *and others* (2001); more details are given in Section 6. As shown in the figure, existing methods each produce estimates of the null component πf_0 that cover one or both tails of the z-score histogram, leaving little to be explained by the nonnull component $(1 - \pi) f_1$. Consequently, zero discoveries of interesting genes are made in one or both tails. Classification-based analyses (e.g., Lee *and others*, 2003), on the other hand, identify interesting genes in both tails for each of the 2 studies (see Section 6).

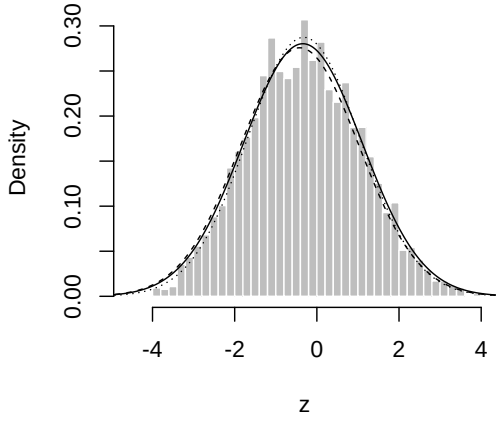
In this paper, we consider a new likelihood-based analysis of the 2-groups model, with a regularization on μ, σ, π, and a semiparametric specification of the nonnull density f_1. We employ a mixture representation of f_1 that gives it heavier tails than f_0 to reflect the belief that z-scores from the nonnull 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. On the other hand, for other studies, where existing methods seem to fail, such as the 2 studies mentioned earlier, our method produces different but arguably more believable results (see Sections 6 and 7). A similar adaptability property manifests in our simulation study in Section 5 where z-scores are generated according to (1.1) with π ranging between 0.75 to 0.99.

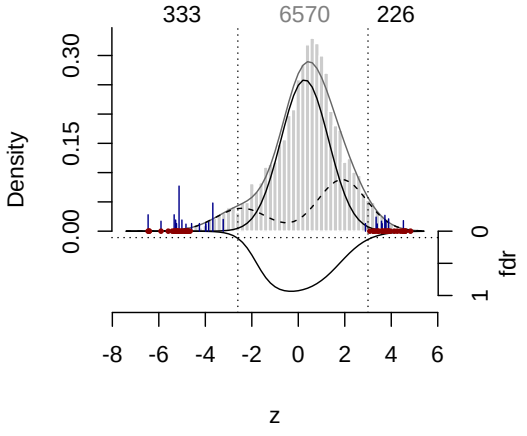
Despite a nonparametric specification of f_1 and a likelihood-based analysis, our treatment of the 2-groups model retains the computational efficiency that is hallmark of methods based on z-scores. This has been possible due to recent developments on a stochastic algorithm due to Newton (2002) called predictive recursion (PR) for estimation of mixing densities with respect to any arbitrary dominating measure (see also Newton *and others*, 1998). Theoretical properties of this algorithm are addressed in Ghosh and Tokdar (2006), Martin and Ghosh (2008), Tokdar *and others* (2009), and Martin and Tokdar (2009). Martin and Tokdar (2011) show how this algorithm can be used in a hierarchical mixture model to



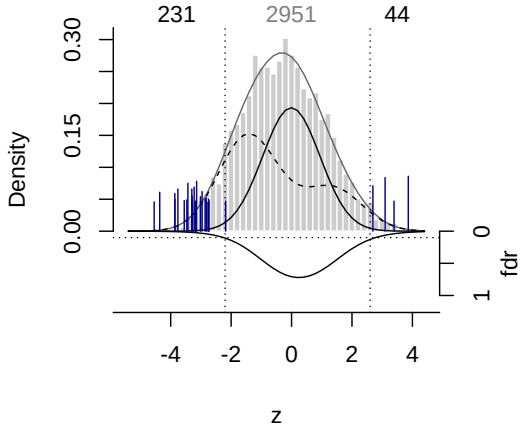
(a) Golub *and others* (1999) z-scores



(b) Hedenfalk *and others* (2001) z-scores



(c) Golub *and others* (1999) z-scores and PRtest fit



(d) Hedenfalk *and others* (2001) z-scores and PRtest fit

Fig. 1. Density histogram of z-scores from leukemia microarray data (Golub *and others*, 1999) and breast cancer data (Hedenfalk *and others*, 2001). Panels (a) and (b) display estimates of πf_0 based on the methods of Efron (2004) (—), Jin and Cai (2007) (− −), and Muralidharan (2010) (· · ·). Panels (c) and (d), discussed in Section 6.2, show the z-score histograms along with the corresponding PRtest fit: πf_0 (solid black line), $(1-\pi) f_1$ (dashed black line), f (solid gray line). Estimated fdr and the 0.1 threshold are shown on the negative scale. The 27 genes identified by Lee *and others* (2003) in each data set are marked with a vertical bar at their z-score; the bar's height represents its posterior inclusion probability in their classification model. For the leukemia data in panel (c), dots point to z-scores of the 50 genes originally identified by Golub *and others* (1999).

construct a likelihood function over nonmixing 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 adapt this marginal likelihood to the 2-groups model, with μ, σ, π, and a scaling parameter in the specification of f_1 serving as the nonmixing 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 (fdr), 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 nonnull density f_1 and to the degree of sparsity characterized by π. Performance of PRtest in an interesting example using the artificial microarray data of Choe *and others* (2005) is addressed in Section 6. In this example, the set of interesting genes is known and we find that PRtest performs considerably better than existing methods and strikingly similar to the oracle. Likewise, for the leukemia and hereditary breast cancer studies, 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 more sophisticated high-dimensional classification-based techniques.

2. MODEL SPECIFICATION

We take $f_0(z) = \mathrm{N}(z|\mu, \sigma^2)$, the normal density with unknown mean and variance μ and σ^2. The nonnull density f_1 is taken to be a semiparametric mixture of the form

$$f_1(z) = \int_{\mathscr{U}} \mathrm{N}(z|\,\mu + \tau\sigma u, \sigma^2)\psi(u)\mathrm{d}u, \tag{2.2}$$

with ψ a density with respect to the Lebesgue measure on $\mathscr{U} = [-1, 1]$ and $\tau \geqslant 1$ a scaling factor. An important consequence of the requirement that ψ be a density is given in the following theorem (for the proof see Appendix A of the supplementary material available at *Biostatistics* online).

THEOREM 2.1 For f_0 and f_1 as described above, the parameters (μ, σ, π, τ, and ψ) in our version of the 2-groups model are identifiable.

This result is useful because, in general, identifiability is not guaranteed for a 2-groups model (1.1) with an empirical null that involves unknown parameters. For our specification, the key to identifiability is the model feature that f_1, by virtue of averaging over locations shifts of f_0, has heavier tails than f_0. This feature is scientifically relevant as it embeds the belief that z-scores in the tails of the histogram are more likely to correspond to nonnull cases than null. Efron (2008) incorporates a similar belief through a "zero-assumption": 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 πf_0 that has heavier tails than any reasonable histogram-smoothing estimate of f, as reported by Strimmer (2008) and illustrated in Figure 1. In comparison, separating f_0 and f_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 f as the mixture model

$$f(z) = \int_{\mathscr{U}} p(z\,|\theta, u)\Psi(\mathrm{d}u) \tag{3.3}$$

with parameters $\theta = (\mu, \sigma, \tau)$, kernel $p(z \mid \theta, u) = \mathrm{N}(z\,|\,\mu + \tau\sigma u, \sigma^2)$, and mixing probability measure Ψ 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 ψ. The collection of all such Ψ is the set $\mathbb{P} = \mathbb{P}(\mathscr{U}, \nu)$ of probability measures that are absolutely continuous with respect to the measure ν defined as the sum of the Lebesgue measure on $\mathscr{U}$ and a point mass at 0. The ν-density of such an Ψ will be denoted by $\pi\langle 0\rangle + (1-\pi)\psi$.

Inference on (θ, Ψ), with Ψ as in (3.3), can be performed in a Bayesian setting with a prior distribution on (θ, Ψ). A popular choice of prior distribution for the nonparametric probability measure Ψ is the Dirichlet process prior (Ferguson, 1973). However, there are 2 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 Ψ 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 can reach several 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 2-groups framework.

As an alternative, we estimate (θ, Ψ) via the PR methodology (Newton, 2002; Martin and Tokdar, 2011). PR is a stochastic algorithm for estimating a mixing distribution Ψ in (3.3) 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; Tokdar *and others*, 2009; Martin and Ghosh, 2008; Martin and Tokdar, 2009). However, Newton's original proposal can estimate the mixing distribution only when the kernel being mixed is known exactly, i.e., for (3.3), an estimate of Ψ is available only when θ is known. To resolve this difficulty, Martin and Tokdar (2011) introduce a "marginal likelihood" function for nonmixing parameters θ based on the output of the PR algorithm.

PR ALGORITHM. Start with an initial estimate Ψ_0 with ν-density $\pi_0\langle 0\rangle + (1-\pi)\psi_0$ and a sequence of weights $w_1, \ldots, w_n \in (0, 1)$. For $i = 1, \ldots, n$, compute

$$f_{i-1,\theta}(Z_i) = \int p(Z_i \,|\, \theta, u)\Psi_{i-1}(\mathrm{d}u),$$

$$\Psi_i(\mathrm{d}u) = (1-w_i)\Psi_{i-1}(\mathrm{d}u) + w_i\, p(Z_i \,|\, \theta, u)\Psi_{i-1}(\mathrm{d}u)/f_{i-1,\theta}(Z_i). \tag{3.4}$$

Produce Ψ_n as an estimate of Ψ and $L_n(\theta) = \prod_{i=1}^n f_{i-1,\theta}(Z_i)$ as the marginal likelihood of θ.

Martin and Tokdar (2011) give 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 Ψ under the Bayesian specification $\Psi \sim \mathrm{DP}(\alpha, \Psi_0)$, the Dirichlet process distribution with precision $\alpha = (1-w_1)/w_1$ and base measure Ψ_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 i.i.d. data models (Wald, 1949). Specifically, for large n, with $Z_1, \ldots, Z_n$ independently drawn from a common density $f^\star$, $\log L_n(\theta) \approx -nK^\star(\theta)$, where $K^\star(\theta)$ equals the minimum Kullback–Leibler divergence between $f^\star$ and densities f of the form (3.3) with Ψ ranging over the set $\mathbb{P}$ and its weak limit points.

4. REGULARIZED PREDICTIVE RECURSION INFERENCE AND PRTEST

We employ a regularized version of the predictive recursion methodology to estimate (θ, Ψ) for our 2-groups model. The regularization is motivated by a hierarchical Bayes formulation of (3.3) with $\Psi \sim \mathrm{DP}(\alpha, \Psi_0)$ where hyper-prior distributions are specified on the model parameters μ, σ, τ, and Ψ_0. We take the ν-density of Ψ_0 to be $\pi_0\langle 0\rangle + (1-\pi_0)\psi_0$ with a fixed choice of $\psi_0(u) \propto u^2$. Among the

remaining parameters, $\sigma \in (0, \infty)$, $\tau \in (1, \infty)$, and $\pi_0 \in (0, 1)$ are taken to be independent with $\log \sigma \sim \mathrm{N}(0, 0.25^2)$, $\log(\tau - 1) \sim \mathrm{N}(0, 1)$, and $\pi_0 \sim \mathrm{Beta}(22.7, 1)$. Given σ and the other parameters, μ is assigned the conditional prior distribution $\mathrm{N}(0, \sigma^2/400)$.

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 nonnull density f_1 is considerably wider than f_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 *and others* (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

$$\tilde{\ell}_n(\mu, \sigma, \tau, \pi_0) = \log L_n(\mu, \sigma, \tau, \pi_0) + \log g(\mu, \sigma, \tau, \pi_0). \tag{4.5}$$

Estimates of these parameters are obtained by maximizing $\tilde{\ell}_n = \tilde{\ell}_n(\mu, \sigma, \tau, \pi_0)$. Once these estimates are obtained, PR is run one last time with the estimated values of these parameters to produce an estimate of F, i.e., of π and of ψ in (1.1) and (2.2), respectively. In our implementations, maximization of ℓ_n is done by the gradient-based Broyden–Fletcher–Goldfarb–Shanno optimization method. Appendix B of the supplementary material available at *Biostatistics* online provides a variation on the PR algorithm that produces the gradient of $\log L_n$ as a by-product.

The PR methodology depends on 2 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 f when the weights are of the form $w_i = (i+1)^{-\gamma}$, $\gamma \in (2/3, 1]$. 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 *and others*, 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 fdr (Efron, 2004), given by

$$\mathrm{fdr}(z) = \pi f_0(z)/f(z),$$

which represents the posterior probability that a case with z-score $Z = z$ is null. Sun and Cai (2007) argue that the local fdr is the fundamental quantity for multiple testing. Once regularized PR estimation of $(\mu, \sigma, \tau, \pi$, and $\psi)$ is completed, a plug-in estimate $\widehat{\mathrm{fdr}}$ of fdr is readily available, and PRtest is implemented by thresholding $\widehat{\mathrm{fdr}}$; that is, we declare case i as nonnull if $\widehat{\mathrm{fdr}}(Z_i) < r$ for some specified threshold $r \in (0, 1)$. According to Efron, this multiple testing rule will control the Benjamini–Hochberg FDR at level r. In our examples, we take $r = 0.1$. This choice, used by Sun and Cai (2007), 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 simulations compared to the benchmark Bayes oracle test that thresholds the true fdr at level 0.1. 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 $f_0(z) = \mathrm{N}(z \mid \mu, \sigma^2)$. Here, we fix $n = 1000$, $\mu = 0$, and $\sigma = 1$. Four choices of f_1 are considered

C1: $f_1(z) = \mathrm{N}(z \mid 0, \sigma^2 + \omega^2)$. Taking $\omega^2 = 13 \approx 2\sigma^2 \log n$ ensures the nonnull 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: $f_1(z) = 0.5 \int_2^4 \mathrm{N}(z \mid u, \sigma^2)\mathrm{d}u$. This choice, used by Muralidharan (2010) and Johnstone and Silverman (2004), exhibits asymmetry and has only slightly heavier tails than the null.

C3: $f_1(z) = 0.67\mathrm{N}(z \mid -3, 2) + 0.33\mathrm{N}(z \mid 3, 2)$. This one is asymmetric and a large portion of its mass is concentrated away from the origin.

C4: $f_1(z) = 0.25 \int_{[-4,-2]\cup[2,4]} \mathrm{N}(z \mid u, \sigma^2)\mathrm{d}u$. This is a symmetrized version of C2. A key feature of this choice is that the unobserved signals are bounded away from zero.

For each of the 4 choices of f_1, we consider 6 choices of π ranging from 0.75 to 0.99, forming a total of 24 simulations settings. Each setting is replicated 500 times and the results are reported below.

Table 1 summarizes the estimates of the null parameters π for each simulation setting. Estimates of (μ, σ) are similarly accurate across methods, models, and sparsity, 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 2 most practically realistic cases, namely C3 and C4, which have smooth nonnull densities with modes on both the left and right side of zero. Also the average computation time for PRtest is roughly 3 s, which compares favorably with that for Jin–Cai (0.7 s) and mixFDR (0.5 s).

Next, we compare the performance of the selected methods based on false nondiscovery rate (FNR), false discovery rate (FDR), power, and Bayes risk. We limit this discussion to nonnull choice C3; 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 procedure. The message is that PRtest is competitive with the other tests in all aspects across a range of sparsity levels. In particular, the 4 tests are similar in terms of FNR for large π, but PRtest is better than mixFDR and Jin–Cai for relatively small π. Also, each of the 4 tests have relatively small FDRs, although the Jin–Cai method has a somewhat unexpected spike, which explains its higher power for large π values. Theoretically, 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 suggests that PRtest may be asymptotically optimal in the sense of Bogdan *and others* (2011).

6. EXAMPLES

6.1 *Validation with spike-in data*

An interesting "spike-in" data set was built by Choe *and others* (2005). The data set itself is artificial—so the set of interesting genes is known—but their careful construction gives it some features of a real control-versus-treatment microarray study. We consider a subset of this data (available in the R package st) consisting of 11 475 genes, of which 1331 are differentially expressed. Z-scores are obtained by taking a Gaussian transform of the standard 2-sample t-test statistics. Figure 3(a) shows histogram of the observed z-scores along with the PRtest fit of the 2-groups mixture model. The estimated density clearly fits

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 4* f_1*'s described in Section* 5

f_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)

the data very well, and the fdr thresholding method flags 235 genes as downregulated. For comparison, Figure 3(b) reports an oracle fit of the 2-groups model, where π is estimated as the known proportion of differentially expressed genes, (μ, σ) are estimated by maximum likelihood based on the null z-scores, and f_1 is estimated by a standard Gaussian kernel estimate based on the nonnull z-scores; the top panel of Table 2 reports the parameter estimates. This oracle procedure is, in some sense, the best fdr thresholding procedure, one can hope for, and it flags 249 genes as downregulated.

For further comparison, we applied the methods of Efron, Jin–Cai, and Muralidharan and the results are summarized in the top panel of Table 2. PRtest and the oracle perform similarly in every respect, while the other methods are substantially different. Only the Jin–Cai method is able to pick out a reasonable set of interesting genes, a bit larger than the sets identified by the oracle and PRtest. However, these additional discoveries result in a 50% increase in FDR.

6.2 *Application to real data*

We applied PRtest, along with the methods of Efron, Jin−Cai and Muralidharan, to the 2 microarray gene expression data sets mentioned in Section 1: the leukemia study by Golub *and others* (1999) and the hereditary breast cancer study by Hedenfalk *and others* (2001). The parameter estimates and gene classifications are summarized in the bottom 2 panels of Table 2. In both data sets, PRtest estimates π to be relatively small and identifies a number of interesting genes, while the others identify none (see Figure 1(c) and (d)). PRtest's findings in these 2 data sets are corroborated by the results of Lee *and others* (2003) who

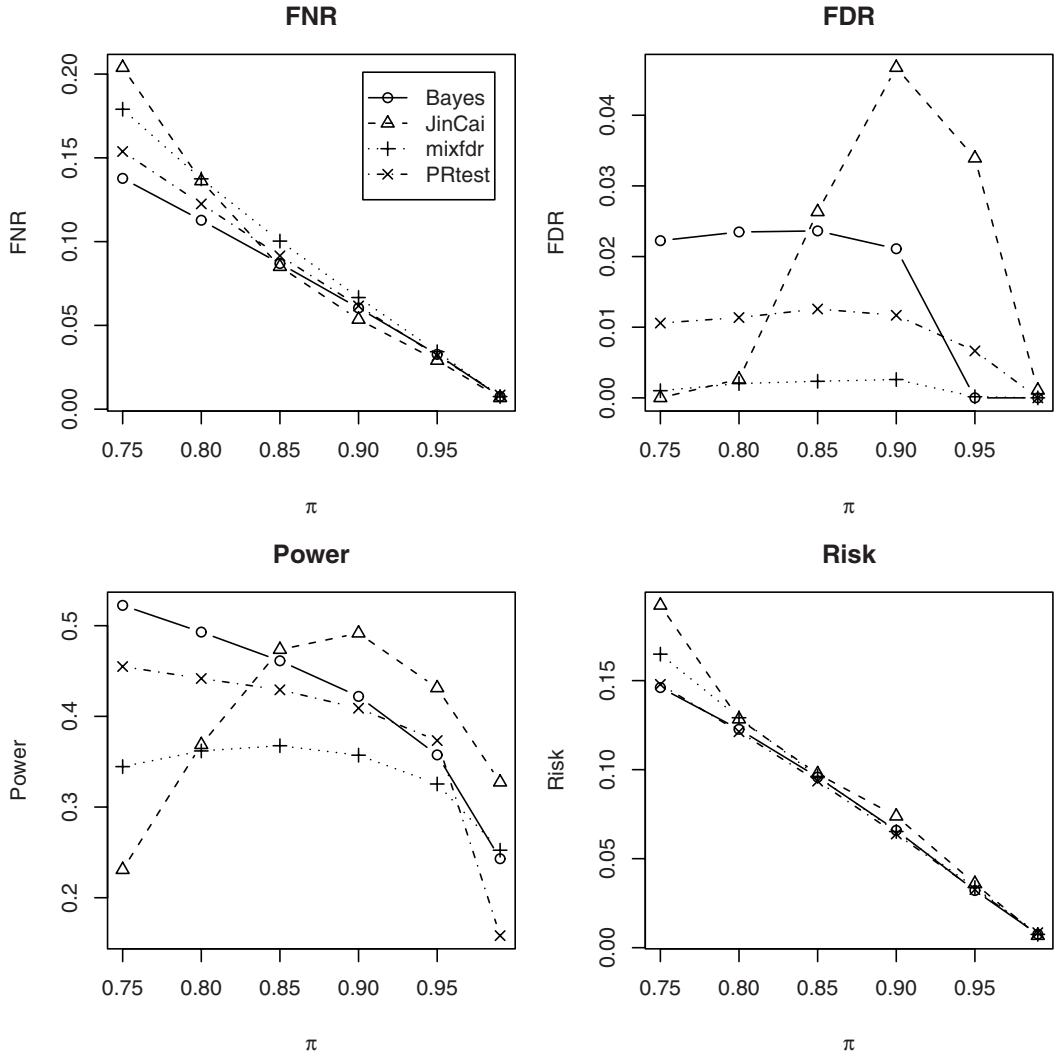
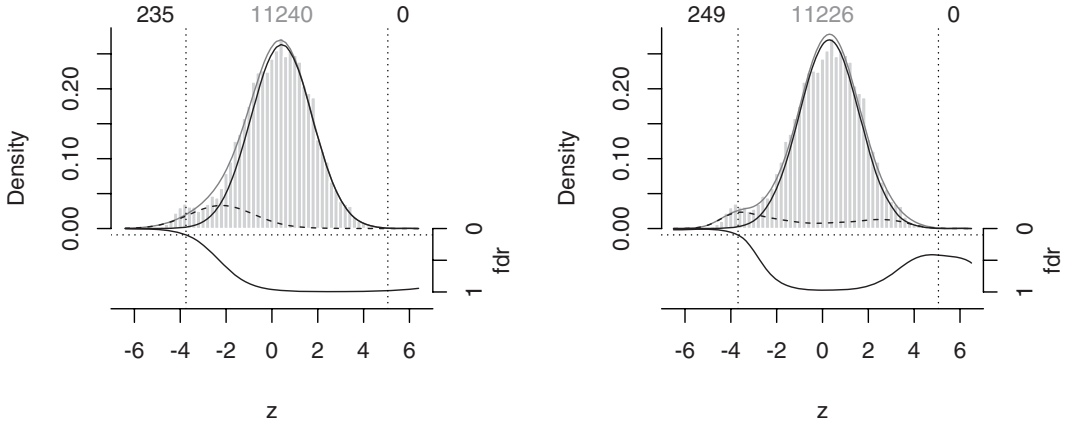


Fig. 2. Plots of the FNR (top left), FDR (top right), power (bottom left), and Bayes risk (bottom right) against π for the selected testing procedures in the C3 simulation setting described in Section 5.

learn a treatment classifier from gene expression levels and validate it by accurately classifying samples from an independent test set. That is, the set of interesting genes identified by PRtest substantially overlaps with the set of genes Lee *and others* (2003) flag as important constituents of their classifier; these are also displayed in Figure 1(c) and (d). For the breast cancer study, some of the genes identified by PRtest and Lee *and others* (2003), such as keratin 8, TOB 1, and phosphofructokinase platelet, have known biological connections to breast cancer mutations (Lee *and others*, 2003, p. 93). The fact that the gene expression levels lead to a well-validated classifier suggests that some genes must be differentially expressed. In this light, it is surprising that the methods of Efron, Jin–Cai, and Muralidharan fail to identify a single interesting gene.



(a) PRtest fit

(b) Oracle fit

Fig. 3. Histogram of the z-scores for the spike-in data in Section 6 along with fits of the 2-groups model using (a) PRtest and (b) the Oracle described in the text. In each plot, overlays are πf_0 (solid black line), $(1-\pi) f_1$ (dashed black line), and f (solid gray line). The estimated fdr and the 0.1 threshold are shown on the negative scale. Numerical values on the top left and right indicate the number of genes flagged as down- and upregulated, respectively, by the fdr thresholding rule.

Table 2. *Results for the 3 data sets considered in Section 6. The "Oracle" method, as described in the text, uses the information about which genes are differentially expressed to estimate fdr*

Data	Method	μ	σ	π		Number of genes		FDR (%)	FNR (%)
						Left	Right		
Spike-in	Efron		0.33	1.50	0.99	2	0	0	12
	Jin–Cai		0.77	1.45	0.91	306	0	3	9
	mixfdr		0.28	1.45	0.97	8	0	0	12
	PRtest		0.42	1.34	0.88	235	0	2	10
	Oracle		0.30	1.31	0.88	249	0	2	10
Leukemia	Efron		0.57	1.18	0.88	276	0	—	—
	Jin–Cai		0.95	1.30	0.91	291	0	—	—
	mixfdr		0.56	1.35	0.96	71	0	—	—
	PRtest		0.23	1.04	0.63	333	226	—	—
BRCA	Efron		−0.33	1.45	1.00	0	0	—	—
	Jin–Cai		−0.42	1.44	1.00	0	0	—	—
	mixfdr		−0.31	1.38	0.99	0	0	—	—
	PRtest		−0.01	1.04	0.45	231	44	—	—

7. Discussion

This paper provides a new and identifiable semiparametric formulation of the 2-groups model and a computationally efficient algorithm to estimate the model parameters. This naturally leads to a nonparametric empirical Bayes multiple testing rule based on thresholding the estimated local fdr. In simulations, we find that PRtest is comparable to existing methods, including the Bayes oracle. What is particularly interesting is that the PRtest results differ substantially from those of existing methods in the examples of Section 6, and we argue that our findings are, in fact, more believable.

We have chosen to focus only on the case where the null z-scores are normally distributed, though the theory and methods presented here work for other well-behaved parametric families. Normality of null z-scores is indeed a strong structural assumption, but identification of the null from the nonnull requires strong parametric shape restrictions on one of the 2 components. Assuming a normal null component is natural because, theoretically, the null z-scores should have a standard normal distribution. This is similar to p-value-based methods where the null p values are assumed to be uniform. A purely statistical verification of this kind of assumption seems quite challenging. One could possibly gain insight on this issue through biological experiments consisting entirely of null cases.

We have justified the continuous location mixture formulation of f_1 in (2.2) on 2 grounds: first, it makes the model parameters identifiable and second it conforms to the accepted notion that the alternative is more likely than the null to produce z-scores of large magnitude. This latter property is also satisfied by a discrete mixture $f_1 = \sum_{j=1}^{J} \pi_j \mathrm{N}(\mu + \tau\sigma u_j, \sigma^2)$ for which the identifiability condition does not hold. But with the regularization to encourage selection of f_0 centered near zero, and the ability of a flexible continuous mixture to approximate a discrete one, PRtest might still perform well in this difficult situation. Our limited simulations seem to indicate that this is true. The case where f_1 is not wider than f_0 also yields a coherent statistical simulation model, but we argue that it corresponds to a biologically untenable abstraction. Indeed, the multiple testing framework accepts the z-scores as scores whose magnitudes (possibly after a small shift of origin) give an ordering of how interesting the cases are relative to each other. The question is to decide how interesting a case must be in order to be labeled as nonnull. Accepting the relative ordering is equivalent to accepting that f_1 must be wider than f_0.

Software

R software for PRtest is available at http://www.stat.duke.edu/~st118/Software.

Supplementary Material

Supplementary material, including a proof of Theorem 2.1 and a recursive algorithm for evaluating the gradient of $\tilde{\ell}_n(\cdot)$ in (4.5), is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors are grateful to the Editor, Associate Editor, and 2 anonymous referees for their insightful comments and suggestions and to Professor J. K. Ghosh for many helpful discussions. A portion of this work was completed while R. Martin was with the Department of Mathematical Sciences, Indiana University–Purdue University Indianapolis. *Conflict of Interest:* None declared.

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.

BOGDAN, M., CHAKRABARTI, A., FROMMLET, F. AND GHOSH, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics* **39**, 1551–1579.

BOGDAN, M., GHOSH, J. K. AND TOKDAR, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In: Balakrishnan, N., Peña, E. and Silvapulle, M. (editors), *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*. Beachwood, OH: IMS, pp. 211–230.

CHOE, S. E., BOUTROS, M., MICHELSON, A. M., CHURCH, G. M. AND HALFON, M. S. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by wholly defined control dataset. *Genome Biology* **6**, R16.

DONOHO, D. L. AND JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probability 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. *Journal of the American Statistical Association* **99**, 96–104.

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

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**, 1–22.

EFRON, B. AND TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70–86.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **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: Fan, J. and Koul, H. (editors), *Frontiers in Statistics*. London: Imperial College 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. *and others* (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. *and others* (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine* **344**, 539–548.

JIN, J. AND CAI, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association* **102**, 495–506.

JOHNSTONE, I. M. AND SILVERMAN, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics* **32**, 1594–1649.

LEE, K. E., SHA, N., DOUGHERTY, E. R., VANNUCCI, M. 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. *Statistical Science* **19**, 265–274.

LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science* **23**, 439–464.

MARTIN, R. AND GHOSH, J. K. (2008). Stochastic approximation and Newton's estimate of a mixing distribution. *Statistical Science* **23**, 365–382.

MARTIN, R. AND TOKDAR, S. T. (2009). Asymptotic properties of predictive recursion: robustness and rate of convergence. *Electronic Journal of Statistics* **3**, 1455–1472.

MARTIN, R. AND TOKDAR, S. T. (2011). Semiparametric inference in mixture models with predictive recursion marginal likelihood. *Biometrika* **98**, 567–582.

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. *The Astronomical Journal* **122**, 3492–3505.

MURALIDHARAN, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics* **4**, 422–438.

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

NEWTON, M. A., QUINTANA, F. A. AND ZHANG, Y. (1998). Nonparametric Bayes methods using predictive updating. In: Dey, D., Müller, P. and Sinha, D. (editors), *Practical Nonparametric and Semiparametric Bayesian Statistics*. Volume 133 of Lecture Notes in Statistics. New York: Springer, pp. 45–61.

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. *The Annals of Applied Statistics* **2**, 153–175.

STOREY, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* **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. *Journal of the American Statistical Association* **102**, 901–912.

TOKDAR, S. T., MARTIN, R., AND GHOSH, J. K. (2009). Consistency of a recursive estimate of mixing distributions. *The Annals of Statistics* **37**, 2502–2522.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics* **20**, 595–601.

[*Received May 2, 2011; revised September 20, 2011; accepted for publication September 26, 2011*]