A nonparametric empirical Bayes framework for large-scale multiple testing[1][2]

Ryan Martin
University of Illinois at Chicago
www.math.uic.edu/~rgmartin

American Statistical Association
Northeastern Illinois Chapter Meeting
10/24/2012

[1] Joint work with Surya T. Tokdar of Duke University.

[2] *Biostatistics* **13**, 427–439, 2012; preprint at arXiv:1106.3885.

Outline

The problem

- In modern problems, researchers are often charged with making inference on many parameters simultaneously.
- A now classical example is the analysis of DNA microarrays:
  - Find a subset of genes which distinguish patients with/without a particular disease, such as diabetes.
  - Expression levels for n genes are measured, goal is to simultaneously test the following sequence of hypotheses:

$$H_{0i} : \text{gene } i \text{ is differentially expressed}, \quad i = 1, \ldots, n.$$

  - In typical microarray applications, $n \sim 10^4$.
- These modern problems are far beyond what classical multiple comparison methods (e.g., Bonferroni, etc) were designed for.

The problem, cont.

- In this large-scale multiple testing problem, there are a number of unique challenges which arise:
  1. The number of tests is massive;
  2. The sample size is relatively small;
  3. The tests are *not* independent.
- A remedy for the first two challenges is what's often referred to as *borrowing strength*.
- The idea is to use data from all cases for each individual test.
  - The now famous Benjamini–Hochberg test is like this;
  - so is hierarchical Bayes.
- My focus is on a **nonparametric empirical Bayes** approach, which is closely tied to semiparametric mixture models.

This talk

- Discuss a general nonparametric empirical Bayes approach to the large-scale multiple testing problem.
- Present the *predictive recursion* (PR) algorithm.
- Define an approximate marginal likelihood.
- Specialize this PR marginal likelihood to the multiple testing problem—a procedure we all **PRtest**.
- Demonstrate the performance of PRtest on real and simulated microarray data.

Outline

Two-groups models

- $Z_1, \ldots, Z_n$ are "z-scores," with Z_i summarizing the expression level of gene i, $i = 1, \ldots, n$.
- The two-groups model assumes

$$Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} f(z) = \pi f_0(z) + (1 - \pi) f_1(z),$$

  with f_0, f_1 the null and non-null densities, and π the null prob.
- This model has a Bayesian flavor, and decisions are typically made via the posterior null probability for each i:
  - Local false discovery rate: $\text{fdr}(z) = \pi f_0(z) \,/\, f(z)$.
  - Reject H_{0i} if $\text{fdr}(Z_i) \leq r$, with $r = 0.1$ say.
- But π, f_0, f typically are *unknown*!

Two-groups model, cont.

- How to find π, f_0, f?
- In most cases, the z-scores are determined in such a way that, in isolation, $Z_i \sim \mathrm{N}(0,1)$ under H_{0i}.
- It is therefore tempting to take $f_0(z) = \mathrm{N}(z \mid 0,1)$ and then estimate π and $f(z)$ with data.
- However, this is often a bad choice (picture below):
  - (Challenge #3) tests are not independent;
  - unobserved covariates;
  - etc...

Empirical null

- Problems 1–3 above can often be remedied by introducing new parameters to be estimated from data.
- This is called the *empirical null*.
- The new two-groups model becomes:

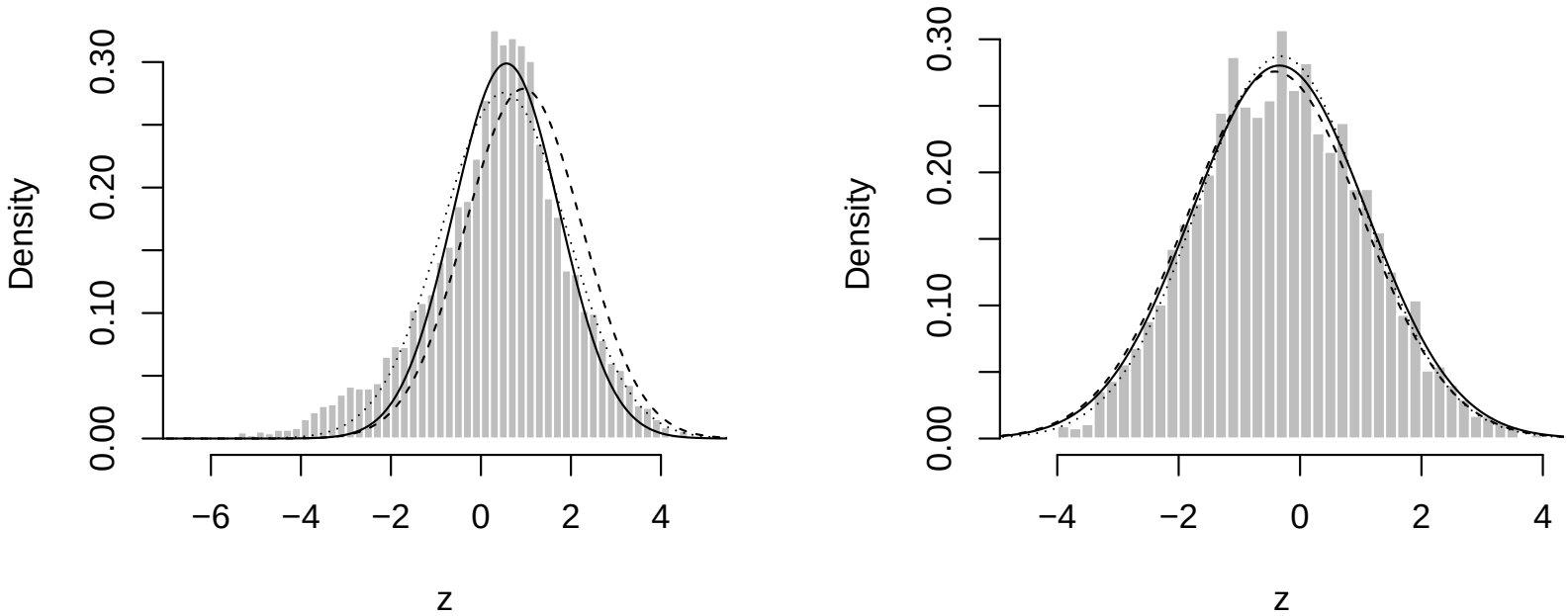
$$f(z) = \pi \mathsf{N}(z \mid \mu, \sigma^2) + (1 - \pi) f_1(z),$$

  where (μ, σ) along with π and $f(z)$ are to be estimated.
- Decisions are made as before, by thresholding the plug-in version $\widehat{\text{fdr}}$ of the local fdr.

Existing implementations

- Two key insights drive the majority of existing methods:
  - Most of the Z_i's are null, so π is large and z-scores carry a lot of information about (μ, σ);
  - n is large, so $f(z)$ can be accurately estimated from data.
- A number of methods exists for this:
  - Efron's locfdr method;
  - Jin and Cai's Fourier-based method;
  - Muralidharan's mixfdr method.
- These methods (and others) ignore f_1 since it's not directly needed to calculate fdr.
- But this neglect of f_1 leads to some strange conclusions in some cases; see figure below.

Two real microarray datasets



(a) Leukemia z-scores

(b) Breast cancer z-scores

Figure: Three estimates of $\pi f_0(z)$.

Outline

Semiparametric mixture model

- Assume $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} f(z)$, modeled as a mixture

$$f_{h,\theta}(z) = \int_{\mathscr{U}} p(z \mid u, \theta) h(u)\, d\lambda(u), \quad h \in \mathbb{H},\ \theta \in \Theta,$$

  where $p(z \mid u, \theta)$ is a known kernel and $\mathbb{H} = \mathbb{H}(\mathscr{U}, \lambda)$ is the set of densities on $\mathscr{U}$ with respect to a measure λ.
- The goal is to estimate the unknown h and θ, although θ is typically of primary importance.

PR algorithm

- PR is a fast stochastic algorithm for estimating mixing densities in nonparametric mixture models.
- Some interesting connections between PR and nonparametric Bayes estimation in Dirichlet process mixture models.
- As PR is only well-defined in the nonparametric mixture model, we shall assume that θ is fixed.
- We'll relax the fixed-θ assumption later.

PR algorithm, cont.

Setup: Suppose $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} f_{h,\theta}$ with θ known.

PR Algorithm.

1. Choose an initial estimate $h_0 \in \mathbb{H}$ and choose a sequence of weights $\{w_i : i \geq 1\} \subset (0, 1)$.
2. For $i = 1, 2, \ldots, n$ compute

$$f_{i-1,\theta}(z) = \int p(z \mid u, \theta) h_{i-1}(u) \, d\lambda(u)$$

$$h_i(u) = (1 - w_i) h_{i-1}(u) + w_i \frac{p(Z_i \mid u, \theta) h_{i-1}(u)}{f_{i-1,\theta}(Z_i)}.$$

3. Return h_n and $f_{n,\theta}$ as estimates of h and $f_{h,\theta}$.

Properties of PR estimates

- Computation is very fast, in particular for low-dim $\mathscr{U}$.
- PR estimates a density with respect to any user-defined dominating measure λ.
  - Maximum likelihood estimate of "h," e.g., is almost surely discrete, regardless of λ.
  - This is important in the testing application to follow!
- Convergence theory helps to choose weights w_i.
- PR estimates depend on the order of the data.
  - Dependence vanishes for large n, usually mild in finite-samples.
  - Averaging h_n over several randomly chosen data permutations reduces the dependence further.

PR convergence theory

- Let $\overline{\mathbb{H}}$ be the weak closure of $\mathbb{H}$.
- Let $K(f, f') = \int \log(f/f')\, f\, dz$ be the KL divergence.
- Define $K^\star(\theta) = \inf\{K(f, f_{h,\theta}) : h \in \overline{\mathbb{H}}\}$.

PR Convergence Theorem.

If $\overline{\mathbb{H}}$ is compact and $p(z \mid u, \theta)$ satisfies some regularity conditions, then $K(f, f_{n,\theta}) \to K^\star(\theta)$ almost surely for each θ.

- If f is identifiable, $K^\star(\theta)$ is attained at a unique $h^\star = h^\star_\theta \in \overline{\mathbb{H}}$, and $h_n \to h^\star$ weakly almost surely.
- What about rates?
  - A general but conservative rate, roughly $o(n^{-1/3})$.
  - For finite-dim $\mathbb{F}$, roughly $o(n^{-1})$.

PR marginal likelihood (PRML)

- How to deal with unknown θ?
- Define the "likelihood function"

$$L_n(\theta) = \prod_{i=1}^{n} f_{i-1,\theta}(Z_i) = \prod_{i=1}^{n} \int p(Z_i \mid u, \theta) h_{i-1}(u)\, d\lambda(u).$$

- Not a bona fide likelihood function for $Z_1, \ldots, Z_n$.
- *However*, $L_n(\theta)$ is a fast approximation to the marginal likelihood in a Dirichlet process mixture model.
- Connection between PRML and Dirichlet process mixture marginal likelihoods also shows up in simulations.
- But what happens to the PRML asymptotically?

PR marginal likelihood, cont.

Define the random function

$$K_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(Z_i)}{f_{i-1,\theta}(Z_i)} = -\frac{\log L_n(\theta) - \sum_{i=1}^{n} \log f(Z_i)}{n}.$$

PRML Convergence Theorem

Under essentially the same conditions of the PR Convergence Theorem, $K_n(\theta) \to K^\star(\theta)$ almost surely for each θ.

- In other words, $\log L_n(\theta) = -nK^\star(\theta) + O(n)$.
- Therefore, maximizing the PRML asymptotically mimicks the oracle procedure of minimizing $K^\star(\theta)$.

Outline

A new two-groups model

Two-groups model

$$
\begin{aligned}
f(z) &= \pi f_0(z) + (1-\pi) f_1(z) \\
&= \pi \mathsf{N}(z \mid \mu, \sigma^2) + (1-\pi) \int_{-1}^{1} \mathsf{N}(z \mid \mu + \sigma\tau u, \sigma^2) h(u)\, du.
\end{aligned}
$$

General semiparametric mixture form, with

$$
\begin{aligned}
\theta &= (\mu, \sigma, \tau), \\
p(z \mid u, \theta) &= \mathsf{N}(z \mid \mu + \sigma\tau u, \sigma^2), \\
\lambda &= \delta_{\{0\}} + \mathsf{Unif}[-1, 1].
\end{aligned}
$$

Properties

- Flexible characterization of the non-null density f_1, rather than leaving it completely unspecified.
- Encodes the belief that f_1 has heavier tails than f_0, i.e., large z-scores are more likely under H_1 than under H_0.

Identifiability Theorem

The new two-groups model described above is identifiable; that is, the map $(\mu, \sigma, \tau, \pi, h) \mapsto f$ is one-to-one.

PRML in this case — PRtest

- PR handles arbitrary base measure λ, so it can estimate a mixing distribution wrt $\lambda = \delta_{\{0\}} + \text{Unif}[-1, 1]$.
- Strategy:
  1. Calculate PRML function as before;
  2. Maximize to estimate (μ, σ, τ);
  3. Run PR once more to estimate (π, h);
  4. Use plug-in estimator of local fdr.
- Computations are relatively fast!
- This two-groups model, with the PRML calculations and fdr thresholding procedure is called **PRtest**.

Outline

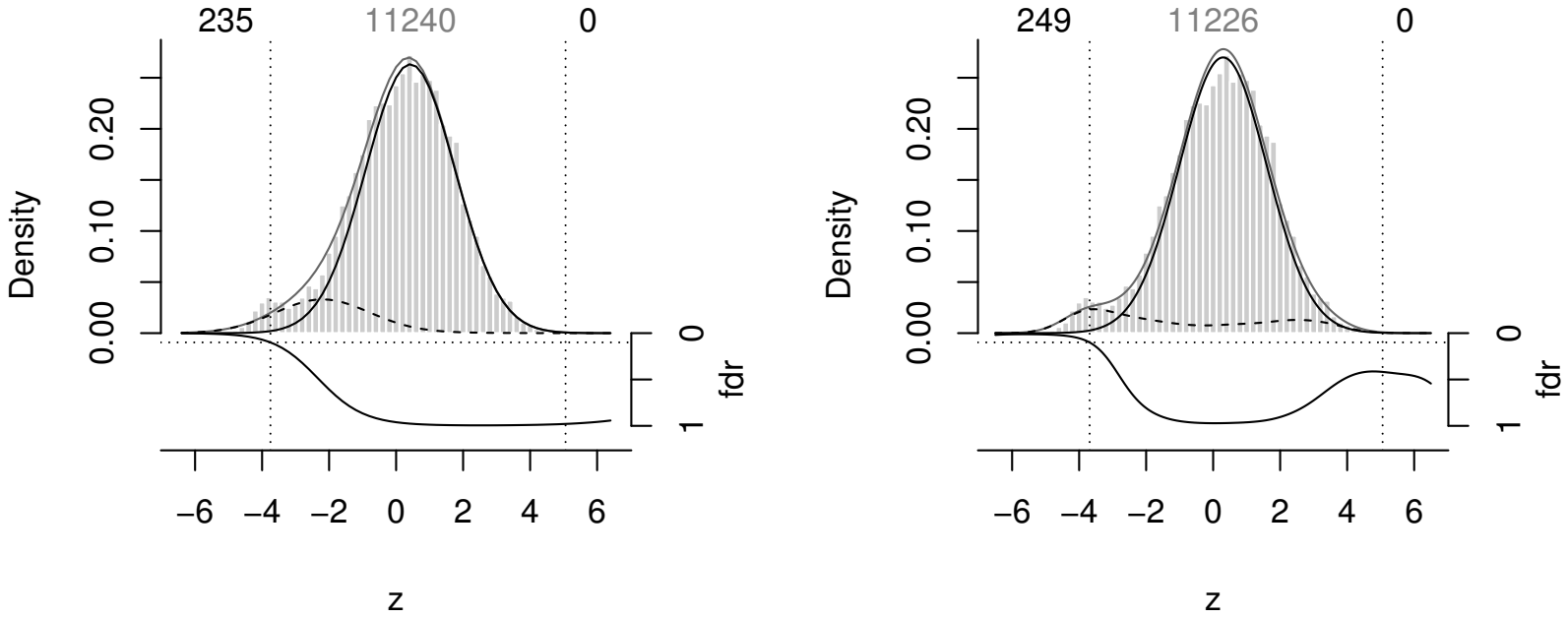
Artificial *spike-in* data

- Choe et al (2005) carefully constructed an artificial microarray data set consisting of $n = 11,475$ genes.
  - Looks a lot like real data;
  - The null and non-null cases are *known*;
  - Useful baseline for comparing different testing procedures.
- Below I compare PRtest with
  - Efron's locfdr;
  - Jin–Cai's Fourier-based method;
  - Muralidharan's mixfdr;
  - *Oracle procedure*

Spike-in data, cont.

Method	μ	σ	π	Number of genes		FDR	FNR
				Left	Right		
locfdr	0.33	1.50	0.99	2	0	0%	12%
Fourier	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%

Spike-in data, cont.



(a) PRtest fit

(b) Oracle fit

Figure: Histogram of the z-scores for the *spike-in* data.

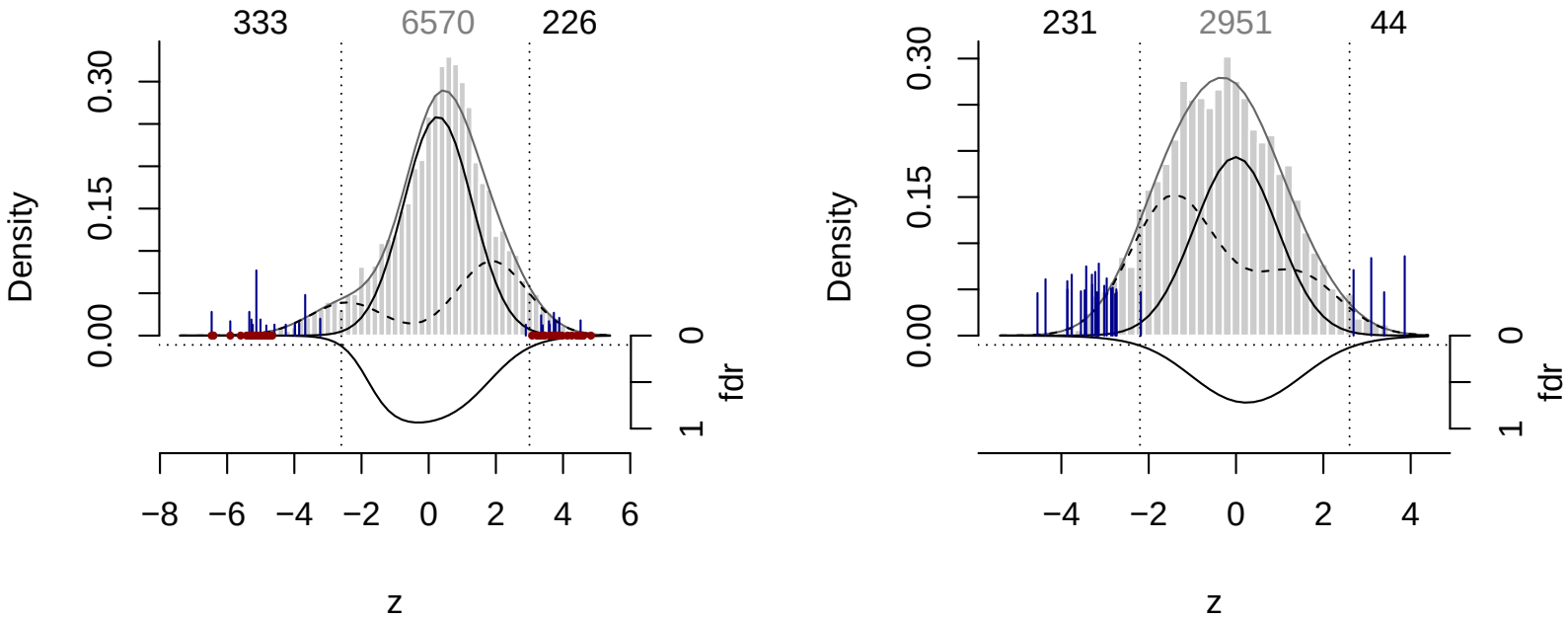
Real microarray data

- Two real microarray data sets:
  - Leukemia (Golub et al)
  - Breast cancer (BRCA; Hedenfalk et al)
- In each, z-scores are obtained by normalizing two-sample t-test statistics for each gene.
- Compare four of the methods from above.

Real microarray data, cont.

Data	Method	μ	σ	π	Number of genes	
					Left	Right
Leukemia	locfdr	0.57	1.18	0.88	276	0
	Fourier	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	locfdr	-0.33	1.45	1.00	0	0
	Fourier	-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

Real microarray data, cont.



(a) Leukemia

(b) Breast cancer

Figure: PRtest's fit to z-scores values in microarray data. Vertical bars denote posterior inclusion probs from Lee et al (*Bioinformatics* 2003)

Outline

Conclusions

- Proposed a new version of the classical two-groups model
  - Flexible specification maintains the tail ordering of f_0, f_1;
  - Model parameters are identifiable.
- PRtest is a computationally efficient procedure for fitting this new two-groups model and carrying out the test.
- PRtest results closely match the oracle in an artificial data set.
- PRtest gives very different results than existing methods in two real data sets.

Discussion

- Theoretical developments:
  - Sharper PR convergence rates?
  - Convergence of $\hat{\theta}_n = \arg\max L_n(\theta)$?
- Methodological developments:
  - PRtest for simultaneous estimation?
  - PRtest for variable selection in regression?

The end

THANK YOU!