False Discovery Rates: A New Deal

Matthew Stephens[1*],
1 Department of Statistics and Department of Human Genetics, University of Chicago, Chicago, IL, USA
∗ **E-mail: Corresponding mstephens@uchicago.edu**

Abstract

Introduction

Since its introduction in 1995 by Benjamini and Hochberg [1], the "False Discovery Rate" (FDR) has quickly established itself as a key concept in modern statistics, and the primary tool by which most practitioners handle large-scale multiple testing in which the goal is to identify the non-zero "effects" among a large number of imprecisely-measured effects.

Here we consider an Empirical Bayes (EB) approach to FDR. This idea is, of course, far from new: indeed, the notion that EB approaches could be helpful in handling multiple comparisons predates introduction of the FDR (e.g. [2]). More recently, EB approaches to the FDR have been extensively studied by several authors, especially B. Efron and co-authors [3–7]; see also [8–11] for example. So what is the "New Deal" here? We introduce two simple ideas that are new (at least compared with existing widely-used FDR pipelines) and can substantially affect inference. The first idea is to *assume that the distribution of effects is unimodal*. This yields a very simple, fast, and stable computer implementation, as well as improving inference of FDR when the unimodal assumption is correct. The second idea is to use two numbers – effect sizes, and their standard errors – rather than just one – p values, or z scores – to summarize each measurement. This idea allows variations in measurement precision to be better accounted for, and avoids a problem with standard pipelines that poor-precision measurements can inflate estimated FDR.

In addition to these two new ideas, we highlight a third idea that is old, but which remains under-used in practice: the idea that it may be preferable to focus on estimation rather than on testing. In principle, Bayesian approaches can naturally unify testing and estimation into a single framework – testing is simply estimation with some positive prior probability that the effect is exactly zero. However, despite ongoing interest in this area from both frequentist [12] and Bayesian [13,14] perspectives, in practice large-scale studies that assess many effects almost invariably focus on testing significance and controlling the FDR, and not on estimation. To help provide a bridge between FDR and estimation we introduce the term "local false sign rate" (lfsr), which is analogous to the "local false discovery rate" (lfdr) [6], but which measures confidence in the *sign* of each effect rather than confidence in each effect being non-zero. We show that in some settings, particularly those with many discoveries, the lfsr and lfdr can be quite different, and emphasise benefits of the lfsr, particularly its increased robustness to modelling assumptions.

Our methods are implemented in an R package, `ashr` (for **a**daptive **sh**rinkage in **R**), available at `http://github.com/stephens999/ashr`. (We address the reasons for this name, and connections with shrinkage analysis, in the Discussion.)

Methods

Model Outine

We begin with the simplest version of our method, and compare it with existing work, before detailing embellishments.

Suppose that we are interested in the values of J "effects" $\beta = (\beta_1, \ldots, \beta_J)$. For example, in a typical genomics application that aims to identify differentially expressed genes, β_j might be the difference in the mean (log) expression of gene j in two conditions. In contexts where FDR methods are applied, interest often focuses on identifying "significant" non-zero effects; that is, in testing the null hypotheses $H_j : \beta_j = 0$. Here we tackle both this problem, and the more general problem of estimating, and assessing uncertainty in, β_j.

Assume that the available data are estimates $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_J)$ of the effects, and corresponding (estimated) standard errors $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_J)$. Our goal is to compute a posterior distribution for β given the observed data $\hat{\beta}, \hat{s}$, which by Bayes theorem can be written as

$$p(\beta|\hat{\beta}, \hat{s}) \propto p(\beta|\hat{s})p(\hat{\beta}|\beta, \hat{s}). \tag{1}$$

For $p(\beta|\hat{s})$ we assume that the β_j are independent from a unimodal distribution g. This unimodal assumption (UA) is a key assumption that distinguishes our approach from previous EB approaches to FDR analysis. A simple way to implement the UA is to assume that g is a mixture of a point mass at 0 and a mixture of *zero-mean* normal distributions:

$$p(\beta|\hat{s}, \pi) = \prod_j g(\beta_j; \pi), \tag{2}$$

$$g(\cdot; \pi) = \pi_0 \delta_0(\cdot) + \sum_{k=1}^{K} \pi_k N(\cdot; 0, \sigma_k^2), \tag{3}$$

where $N(\cdot; \mu, \sigma^2)$ denotes the density of the normal distribution with mean μ and variance σ^2. Here the mixture proportions $\pi = (\pi_0, \ldots, \pi_K)$ are hyperparameters, which are non-negative and sum to one, and are to be estimated, while the mixture component standard deviations $\sigma_1, \ldots, \sigma_K$ represent a large and dense grid of *fixed* positive numbers spanning a range from very small to very big (so K is fixed and large). (We encourate the reader to think of this grid as becoming infinitely large and dense, as a non-parametric limit, although of course in practice we use a finite grid – see Implementation Details.)

For the likelihood $p(\hat{\beta}|\beta, \hat{s})$ we assume

$$p(\hat{\beta}|\beta, \hat{s}) = \prod_j N(\hat{\beta}_j; \beta_j, \hat{s}_j^2). \tag{4}$$

Here, in additional to some conditional independence assumptions, we are effectively assuming that the number of observations used to compute $\hat{\beta}_j, \hat{s}_j$ are sufficiently large to justify a normal approximation.

This simple model features both the key ideas we want to emphasise in this paper: the UA is encapsulated in (3) while the different measurement precision of different observations is

encapsulated in the likelihood (4) – specifically, observations with larger standard error will have a flatter likelihood, and therefore have less impact on inference. However, this simple model also has several additional assumptions that can be relaxed. Specifically,

1. The use of a mixture of zero-mean normals (3) also implies that g is symmetric about 0; more flexibility can be obtained by replacing the mixture of normals with mixtures of uniforms [see (10)].

2. The model (2) assumes that the effects are identically distributed, independent of their standard errors $\hat{s}$. We can relax this to allow for a relationship between these quantities [see (11)].

3. The likelihood (4) effectively assumes that the number of observations used to compute $\hat{\beta}_j, \hat{s}_j$ are sufficiently large to justify a normal approximation. We can generalize this likelihood using a t likelihood [see (12)].

Of course there remain limitations that are harder to relax, most notably the independence and conditional independence assumptions encapsulated in our model (which are also made by most existing EB approaches to this problem). Correlations among tests certainly arise in practice, either due to genuine correlations in the system of study, or due to unmeasured confounders, and their potential to impact results of an FDR analysis is important to consider whatever analysis methods are used: see [**?**, **?**] for relevant discussion.

Fitting the model

In words, the model above assumes that the effects β_j are independent and identically distributed from a mixture of zero-centered normal distributions, and each observation $\hat{\beta}_j$ is a noisy measurement of β_j with standard error $\hat{s}_j$. Together, these assumptions imply that the observations $\hat{\beta}_j$ are also independent observations, each from a mixture of normal distributions:

$$p(\hat{\beta}|\hat{s}, \pi) = \prod_j [\sum_{k=0}^{K} \pi_k N(\hat{\beta}_j; 0, \sigma_k^2 + \hat{s}_j^2)], \tag{5}$$

where we define $\sigma_0 := 0$.

The usual EB approach to fitting this model would involve two simple steps:

1. Estimate the hyperparameters π by maximizing the likelihood $L(\pi)$, given by (5), yielding $\hat{\pi} := \arg\max L(\pi)$.

2. Compute quantities of interest from the conditional distributions $p(\beta_j|\hat{\beta}, \hat{s}, \hat{\pi})$. For example, the evidence against the null hypothesis $\beta_j = 0$ can be summarized by $p(\beta_j \neq 0||\hat{\beta}, \hat{s}, \hat{\pi})$.

Both steps 1 and 2 are very straightforward: $\hat{\pi}$ can be obtained using a simple EM algorithm [16], and the conditional distributions $p(\beta_j|\hat{\beta}_j, \hat{s}_j, \hat{\pi})$ are analytically available, each a mixture of a point mass on zero and K normal distributions. (Note that the simplicity of the EM algorithm

in step 1 is due to our using a fixed grid for σ_k in (3), instead of estimating σ_k, which may seem more natural but is not straightforward when $\hat{s}_j$ varies among j. This simple device may be useful in other applications.)

In practice we make a small modification to this procedure: instead of obtaining $\hat{\pi}$ by maximizing the likelihood, we instead maximise a penalized likelihood [see (15)], where the penalty encourages $\hat{\pi}_0$ to be as big as possible whilst remaining consistent with the observed data. We introduce this penalty because in FDR appllications it is considered desirable to avoid underestimating π_0 so as to avoid underestimating the FDR.

The local False Discovery Rate and local False Sign Rate

As noted above, the posterior distributions $p(\beta_j|\hat{\beta}, \hat{s}, \hat{\pi})$ have a simple analytic form. In practice it is common, and desirable, to summarize these distributions to convey the "significance" of each observation j. One natural measure of the significance of observation j is its "local FDR" [6], which is the probability, given the observed data, that effect j would be a false discovery, if we were to declare it a discovery. In other words it is the posterior probability that β_j is actually zero:

$$\text{lfdr}_j := \Pr(\beta_j = 0|\hat{\beta}, \hat{s}, \hat{\pi}). \tag{6}$$

The lfdr, like most other measures of significance (e.g. p values and q values), is rooted in the hypothesis testing paradigm which focuses on whether or not an effect is exactly zero. This paradigm is popular, despite the fact that many statistical practitioners have argued that it is often inappropriate because the null hypothesis $H_j : \beta_j = 0$ is often implausible. For example, Tukey ([17]) argued that "All we know about the world teaches us that the effects of A and B are always different – in some decimal place – for any A and B. Thus asking 'Are the effects different?' is foolish." Instead, Tukey suggested ([18], p32,) that one should address

> ...the more meaningful question: "is the evidence strong enough to support a belief that the observed difference has the correct sign?"

Along the same lines, Gelman and co-authors [14, 19] suggest focussing on "type S errors", meaning errors in sign, rather than the more traditional type I errors.

Motivated by these suggestions, we define the "local False Sign Rate" for effect j, lfsr_j, to be the probability that we would make an error in the sign of effect j if we were forced to declare it either positive or negative. Specifically,

$$\text{lfsr}_j := \min[p(\beta_j \geq 0|\hat{\beta}, s), p(\beta_j \leq 0|\hat{\pi}, \hat{\beta}, s)]. \tag{7}$$

To illustrate, suppose for concreteness that $p(\beta_j < 0|\hat{\beta}, s, \hat{\pi})) = 0.95, p(\beta_j = 0|\hat{\beta}, s, \hat{\pi})) = 0.03, p(\beta_j > 0|\hat{\beta}, s, \hat{\pi}) = 0.02$. Then from (7) $\text{lfsr}_j = \min(0.05, 0.98) = 0.05$ (and, from (6), $\text{lfdr}_j = 0.03$). This lfsr corresponds to the fact that, given these results, our best guess for the sign of β_j is that it is negative, and the probability that this guess is wrong would be 0.05.

As our notation suggests, lfsr_j is intended to be compared and contrasted with lfdr_j: whereas small values of lfdr_j indicate that we can be *confident that β_j is non-zero*, small values of lfsr_j indicate that we can be *confident in the sign of β_j*. Of coure, being confident in the sign of an effect logically implies that we are confident it is non-zero, and this is reflected in the fact that

$\text{lfsr}_j \geq \text{lfdr}_j$ (this follows from the definition because both the events $\beta_j \geq 0$ and $\beta_j \leq 0$ in (7) include the event $\beta_j = 0$). In this sense, as a measure of "significance", lfsr is more conservative than lfdr. More importantly, as we illustrate in Section **??**, lfsr can be substantially more robust to modelling assumptions than lfdr.

From these "local" measures of significance, we can also compute average error rates over subsets of observations $\Gamma \subset \{1, \ldots, J\}$. For example,

$$\widehat{\text{FDR}}(\Gamma) := (1/|\Gamma|) \sum_{j \in \Gamma} \text{lfdr}_j. \tag{8}$$

estimates the FDR we would obtain if we were to declare all tests in Γ significant. And

$$q_j := \widehat{\text{FDR}}(\{k : \text{lfdr}_k \leq \text{lfdr}_j\}) \tag{9}$$

provides a measure of significance analogous to Storey's q value [20].

See Appendix for a more thorough discussion of the connection between the Bayesian quantities we compute here and frequentist quantities, including a brief discussion of the False Sign Rate from a frequentist perspective.

Related work

Previous approaches focussed on FDR

Among previous methods that explicitly consider the estimation of False Discovery Rates and related quantities, our work here seems most naturally compared with the Empirical Bayes methods of [6] and [11] (implemented in the R packages `locfdr` and `mixfdr` respectively) and with the widely-used methods from [20] (implemented in the R package `qvalue`), which although not formally an EB approach, shares some elements in common.

There are two key differences between the approach implemented here and all of these three existing methods. First, whereas these existing methods summarize the information on β_j by a single number – either a z score (`locfdr` and `mixfdr`), or a p value (`qvalue`) – we instead work with two numbers $(\hat{\beta}_j, \hat{s}_j)$. Here we are building on [21], who develops Bayesian tests (Bayes Factors) for individual null hypotheses using these two numbers, using the normal approximation 4. Using two numbers instead of one clearly has the potential to be more informative, and indeed, as an example of this, Section **??** illustrates how it can improve performance by taking better account of variation in measurement precision among observations. (And in the special case where $\hat{s}_j$ are equal for all j, our approach essentially comes down to the same thing as these existing methods.)

Second, our unimodal assumption (UA) that the effects are unimodal about zero is quite different from assumptions made by `qvalue`, `locfdr` or `mixfdr`. Indeed, `locfdr` assumes that all z scores near 0 are null (Efron calls this the Zero Assumption; ZA), which implies that under the alternative hypothesis the distribution of z scores has *no mass at 0*; this contrasts strikingly with the UA, which implies that this distribution has its peak at 0! Similarly, `qvalue` assumes that all p values near 1 are null, which is the same as the ZA because p values near 1 correspond to z scores near 0. And although `mixfdr` does not formally make the ZA, we have found that in practice, with default settings, the results often approximately satisfy the ZA (due, we believe,

to the default choice of penalty term β described in [11]). Thus, not only do these existing methods not make the UA, they actually make assumptions that are, in some sense, as different from the UA as they can be.

Given that the UA and ZA are so different, it seems worth discussing why we favor the UA. Although the UA will not apply to all situations, we believe that it will often be reasonable, especially in FDR-related contexts that have traditionally focussed on rejecting the null hypotheses $\beta_j = 0$. This is because if "$\beta_j = 0$" is a plausible null hypothesis, it seems reasonable to expect that "β_j very near 0" is also plausible. Further, it seems reasonable to expect that larger effects become decreasingly plausible, and so the distribution of the effects will be unimodal about 0. To paraphrase Tukey, "All we know about the world teaches us that large effects are rare, whereas small effects abound." We emphasise that the UA relates to the distribution of *all* effects, and not only the *detectable* effects (i.e. those that are significantly different from zero). It is very likely that the distribution of *detectable* non-zero effects will be multimodal, with one mode for detectable positive effects and another for detectable negative effects, and the UA does not contradict this.

Alternatively, we could motivate the UA by its effect on point estimates, which is to "shrink" the estimates towards the mode - such shrinkage is desirable from several standpoints for improving estimation accuracy. Indeed most model-based approaches to shrinkage make parametric assumptions that obey the UA (e.g. [22]). Further, almost all analogous work in sparse regression models make the UA for the regression coefficents - common choices of uni-modal distribution being the spike and slab, Laplace, t, normal-gamma, normal-inverse-gamma, or horseshoe priors [23]. These are all less flexible than the approach we take here, which provides for general uni-modal distributions, and it may be fruitful to apply our methods to the regression context; indeed see [24] for work in this vein. The UA assumption on regression coefficients is directly analagous to our UA for effects, and so we view its widespread use in the regression context as supporting its use here.

In addition to these arguments, the UA also has a considerable practical benefit: it produces very simple procedures that are both computationally and statistically stable. We illustrate these features in Results.

Other work

There is also a very considerable literature that does not directly focus on the FDR problem, but which involves essentially the same models, and deals with closely-related issues. Among these, [25] is perhaps most similar to our work: indeed all the elements of our approach outlined above, except for the point mass on 0 and corresponding penalty term, appear in their paper. However, they focus entirely on the estimation of g, in contrast to our focus on estimating β_j; and they provide no software implementation. More generally, the literature is too large to do a comprehensive review here, but relevant key-words include "empirical bayes", "shrinkage", "deconvolution", "semi-parametric", "shape-constrained", and "heteroskedastic". Some pointers to recent papers in which other relevant citations can be found include [26–28]. Much (but not all) of the literature focusses on the homoskedastic case (i.e. $\hat{s}_j$ all equal) whereas we allow for heteroskedasticity. And much (but not all) of the recent shrinkage-oriented literature focuses only on point estimation of β_j, whereas for FDR-related applications measures of uncertainty are

essential. Several recent papers consider more flexible non-parametric assumptions on g than the UA assumption we make here. In particular, [28, 29] consider the unconstrained non-parametric maximum likelihood estimate (NPMLE) for g. These methods may provide alternatives to our approach in settings where the UA assumption is considered too restrictive. However, the NPMLE for g is a discrete distribution, which will induce a discrete posterior distribution on β_j, and so although the NPMLE may perform well for point estimation, it seems possible it will not adequately reflect uncertainty in β_j, and some regularization on g may be necessary.

Embellishments

More flexible unimodal distributions

Using a mixture of zero-centered normal distributions for g in (3) implies that g is not only unimodal, but also symmetric. Furthermore, even some symmetric unimodal distributions, such as those with a flat top, cannot be well approximated by a mixture of zero-centered normals. Therefore, we have implemented a more general approach, based on

$$g(\cdot;\pi) = \sum_{k=0}^{K} \pi_k f_k(\cdot) \tag{10}$$

where f_0 is a point mass on 0, and f_k $(k = 1, \ldots, K)$ are pre-specified component distributions with one of the following forms:

(i) $f_k(\cdot) = N(\cdot; 0, \sigma_k^2)$,

(ii) $f_k(\cdot) = U[\cdot; -a_k, a_k]$,

(iii) $f_k(\cdot) = U[\cdot; -a_k, 0]$ and/or $U[\cdot; 0, a_k]$,

where $U[\cdot; a, b]$ denotes the density of a uniform distribution on $[a, b]$. (In (iii) we include both components in the mixture (10), so a grid of values $a_1, \ldots, a_K$ defines $2K+1$ mixture component densities, and π is a $2K + 1$ vector that sums to 1.) The simplest version (3) corresponds to (i). Replacing these with uniform components (ii)-(iii) only slightly complicates calculations under the normal likelihood (4), and greatly simplifies the calculations under the t likelihood (12) introduced below. The use of uniform components here closely mirrors [25]. (In fact our implementation can handle *any* prespecified uniform or normal distributions for f_k provided they are all from the same family; however, we restrict our attention here to (i)-(iii) which imply a unimodal g.)

Moving from (i) to (iii) the representation (10) becomes increasingly flexible. Indeed, using a large dense grid of σ_k^2 or a_k, we can allow g to approximate, with arbitrary accuracy,

(i) any scale mixture of normals, which includes as special cases the double exponential (Laplace) distribution, any t distribution, and a very large number of other distributions used in high-dimensional regression settings.

(ii) any symmetric unimodal distribution about 0.

(iii) any unimodal distribution about 0.

The latter two claims are related to characterizations of unimodal distributions due to [30] and [31]; see [32], p158. In other words, (ii) and (iii) provide fully non-parametric estimation for g under the constraints that it is (ii) both unimodal and symmetric, or (iii) unimodal only.

Although our discussion above emphasises the use of large K, in practice modest values of K can provide reasonable performance. The key point is that the value of K is not critical provided it is sufficiently large, and the grid of σ_k or a_k values suitably chosen. See Section 0.1 for our software defaults.

Dependence of effects on standard errors

Equation (**??**) assumes that the β_j all come from the same distribution g, independent of $\hat{s}_j$. This can be relaxed to allow the distribution of β_j to depend on $\hat{s}_j$ using

$$\frac{\beta_j}{\hat{s}_j^\alpha} \,|\hat{s}_j \sim g(\cdot;\pi) \tag{11}$$

for any $\alpha \geq 0$. Setting $\alpha = 0$ yields (**??**), and setting $\alpha = 1$ corresponds to assuming that the $t_j = \beta_j/\hat{s}_j$ have a common distribution. This case is of special interest: it effectively corresponds to the "p value prior" in [21] and is, implicitly, the assumption made by existing FDR methods that rank tests by their p values (or z or t scores). See Results for further discussion.

The model 11 for general α can be fitted using the algorithm for $\alpha = 0$. To see this, define $b_j := \beta_j/\hat{s}_j^\alpha$, and $\hat{b} := \hat{\beta}_j/\hat{s}_j^\alpha$. Then $\hat{b}_j$ is an estimate of b_j with standard error $\hat{s}'_j := \hat{s}_j^{1-\alpha}$. Applying the algorithm for $\alpha = 0$ to effect estimates $\hat{b}_1, \ldots, \hat{b}_J$ with standard errors $\hat{s}'_1, \ldots, \hat{s}'_J$ yields a posterior distribution $p(b_j|\hat{s}_j, \hat{b}_j, \hat{\pi}, \alpha)$, which induces a posterior distribution on $\beta_j = b_j\hat{s}_j^\alpha$.

Different values of α in (11) essentially correspond to different models for how effect sizes β_j scale with the standard error $\hat{s}_j$. Since this scaling is unknown it may be desirable to estimate it from the data, and this can be done here by jointly maximizing the likelihood $p(\hat{b}|\hat{s}, \pi, \alpha)$ over π and α rather than just π. Although values of α other than 0 and 1 may not be easy interpret, and the actual relationship between β_j and $\hat{s}_j$ is unlikely to exactly follow (11) for any α, this approach seems preferable in practice than simply assuming $\alpha = 0$ or $\alpha = 1$. We do not investigate this in detail here, but it has been implemented in our software by C.Dai, using a simple grid search over α.

Replace normal likelihood with t likelihood

We generalize the normal likelihood (4) by replacing it with a t likelihood:

$$\hat{\beta}_j \,|\, \beta_j, \hat{s}_j \sim T_\nu(\beta_j, \hat{s}_j) \tag{12}$$

where $T_\nu(\beta_j, \hat{s}_j)$ denotes the distribution of $\beta_j + \hat{s}_j T_\nu$ where T_ν has a standard t distribution on ν degrees of freedom, and ν denotes the degrees of freedom used to estimate $\hat{s}_j$ (assumed known, and for simplicity assumed to be the same for each j). The normal approximation (4) corresponds to the limit $\nu \to \infty$. This generalization does not complicate inference when the mixture components f_k in (10) are uniforms. When the f_k are normal the computations with a t likelihood are considerably more difficult and we have not implemented this combination.

Equation (12) is, of course, motivated by the standard asymptotic result

$$(\hat{\beta}_j - \beta_j)/\hat{s}_j \sim T_\nu. \tag{13}$$

However (13) does not imply (12), because in (13) $\hat{s}_j$ is random whereas in (12) it is conditioned on. In principle it would be preferable, for a number of reasons, to model the randomness in $\hat{s}_j$; we are currently pursuing this improved approach in work with Mengyin Lu, and results will be published elsewhere. We note here that modelling the randomness in $\hat{s}_j$ is particularly helpful when the number of observations used to estimate $\hat{s}_j$ is very small (e.g. < 10), where our simpler approach can produce unsatisfactory results (see later).

Non-zero mode

An addition to our software implementation, due to C.Dai, allows the mode to be estimated from the data by maximum likelihood.

Implementation Details

See Section 0.1 for further implementation details.

Results

We compare results of `ashr` with existing FDR-based methods implemented in the R packages `qvalue` (v1.99.1 from Bioconductor), `locfdr` (v1.1-7 from `https://cran.r-project.org/src/contrib/Archive/locfdr/`), and `mixfdr` (v1.0, from `https://cran.r-project.org/src/contrib/Archive/mixfdr/`). In all our simulations we assume that the test statistics follow the expected theoretical distribution under the null, and we indicate this to `locfdr` using `nulltype=0` and to `mixfdr` using `theonull=TRUE`. Otherwise all packages were used with default options.

Effects of the Unimodal Assumption

Here we consider the effects of making the UA. In this section, to isolate these effects we consider the simplest case, where every observation has the same standard error, $s_j = 1$ and all methods are provided that information. That is, $\hat{\beta}_j|\beta_j \sim N(\beta_j, 1)$ and $\hat{s}_j = s_j = 1$. In this case the z scores $z_j := \hat{\beta}_j/\hat{s}_j = \hat{\beta}_j$, so modelling the z scores is the same as modelling the $\hat{\beta}_j$, and so the only difference between our method and methods like `locfdr` and `mixfdr` are in how they estimate g.

To briefly summarize the results in this section:

1. The UA can produce very different inferences compared with the ZA.
2. The UA can yield conservative estimates of the proportion of true nulls, π_0, and hence conservative estimates of lfdrand FDR.
3. The UA results in a stable procedure, both numerically and statistically, and is somewhat robust to deviations from unimodality.

The UA and ZA can produce different inferences

To illustrate the different inferences from the UA and ZA we show results for a single dataset simulated with the true effects $\beta_j \sim N(0,1)$ (so with $s_j = 1$, $\hat{\beta}_j \sim N(0, 2^2)$). We used each of the methods `qvalue`, `locfdr`, `mixfdr` and `ashr` to decompose the z scores ($z_j = \hat{\beta}_j$), or their corresponding p values, into null and alternative components. The results (Figure 1) illustrate the clear difference between the existing methods and our method. The effects of the ZA made by `qvalue` and `locfdr` are visually clear, producing a "hole" in the alternative z score distribution around 0. Although `mixfdr` does not formally make the ZA, its decomposition exhibits a similar hole. In contrast, due to the UA, the alternative z score distribution for `ashr` is required to have a mode at 0, effectively "filling in" the hole. (Of course the null distribution also has a peak at 0, and the local fdr under the UA is still smallest for z scores that are far from zero – i.e. large z scores remain the "most significant".)

Figure 1 may also be helpful in understanding the interacting role of the UA and the penalty term (15) that attempts to make π_0 as "large as possible" while remaining consistent with the UA. Specifically, consider the panel of Figure 1 that shows `ashr`'s decomposition of z scores, and imagine increasing π_0 further. This would increase the null component (dark blue) at the expense of the alternative component (light blue). Because the null component is $N(0,1)$, and so is biggest at 0, this would eventually create a "dip" in the light-blue histogram at 0. The role of the penalty term is to push the dark blue component as far as possible, right up to (or, to be conservative, just past) the point where this dip appears. In contrast the ZA pushes the dark blue component until the light-blue component *disappears* at 0. See `https://stephens999.shinyapps.io/unimodal/unimodal.Rmd` for an interactive demonstration.

The UA can produce conservative estimates of π_0

The illustrative example in Figure 1 suggests that the UA will produce smaller estimates of π_0 than the ZA. Consequently `ashr` will estimate smaller lfdrs and FDRs than existing methods that make the ZA. This is desirable, provided that these estimates remain conservative: that is, that π_0 does not underestimate the true π_0 and lfdr does not underestimate the true lfdr. The penalty term (15) aims to ensure this conservative behaviour. To check its effectiveness we performed simulations under various alternative scenarios (i.e. various distributions for the non-zero effects, g_1), and values for π_0. The alternative distributions are shown in Figure 2a, with details in Table 2. They range from a "spiky" distribution – where many non-zero β are too close to zero to be reliably detected, making reliable estimation of π_0 essentially impossible – to a much flatter distribution, which is a normal distribution with large variance ("big-normal") – where most non-zero β are easily detected making reliable estimation of π_0 easier. We also include one asymmetric distribution ("skew"), and one clearly bimodal distribution ("bimodal"), which, although we view as generally unrealistic, we include to assess robustness of `ashr` to deviations from the UA.

For each simulation scenario we simulated 100 independent data sets, each with $J = 1000$ observations. For each data set we simulated data as follows:

1. Simulate $\pi_0 \sim U[0,1]$.
2. For $j = 1, \ldots, J$, simulate $\beta_j \sim \pi_0 \delta_0 + (1 - \pi_0) g_1(\cdot)$.

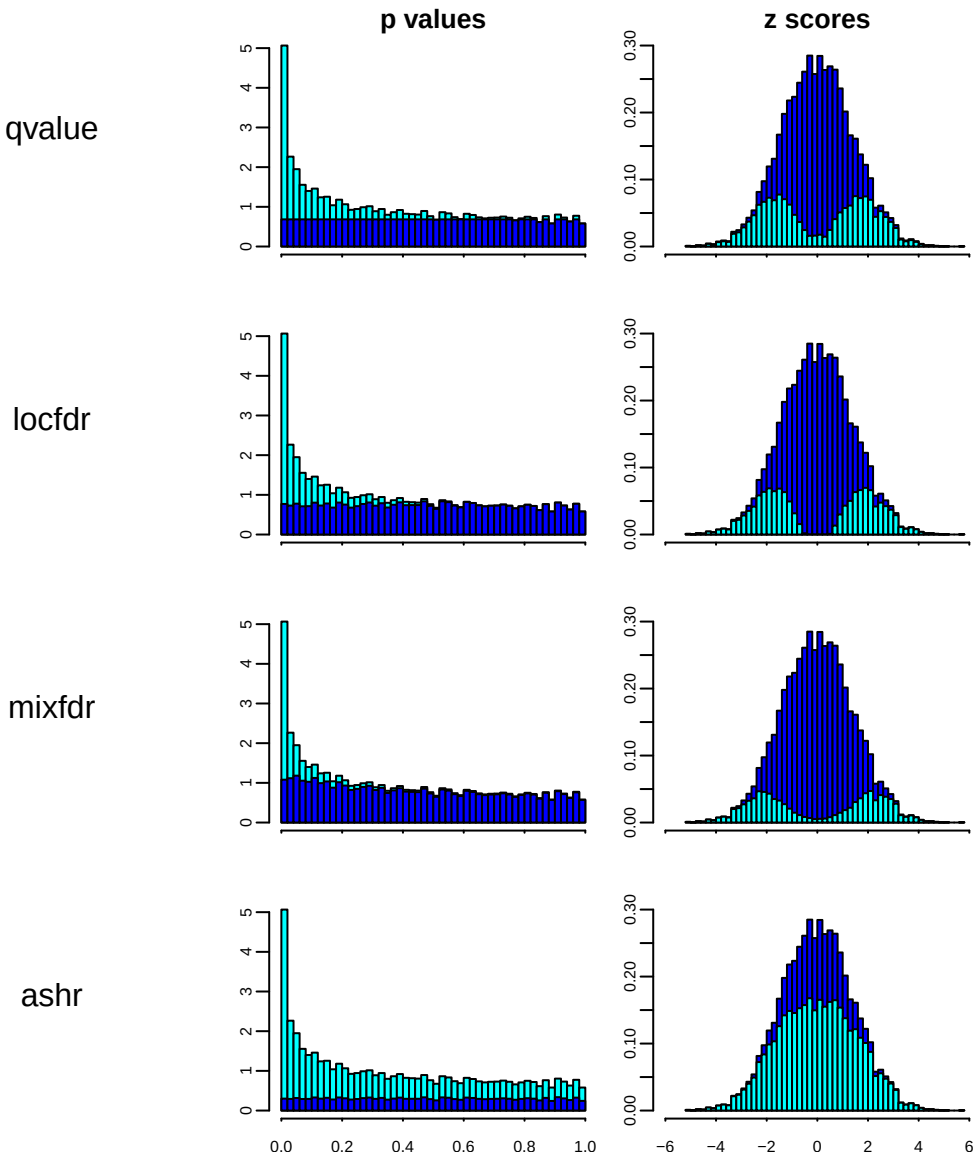
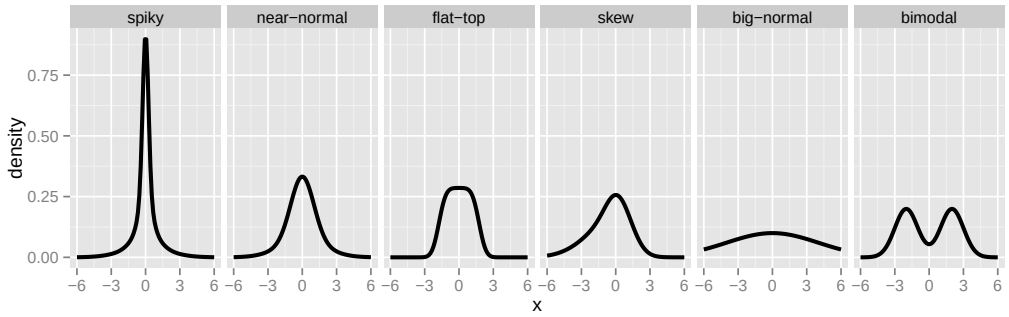
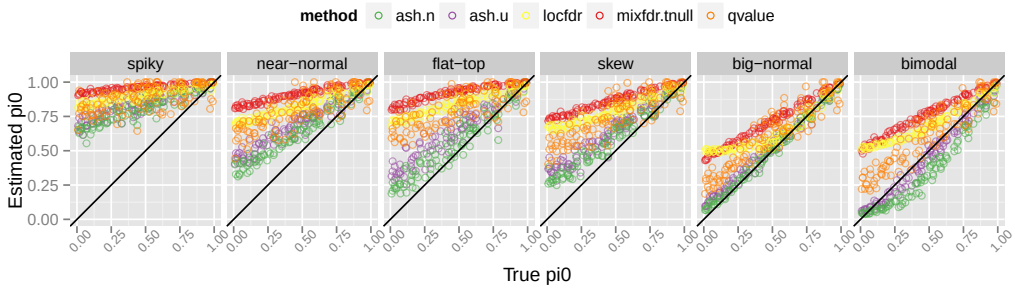


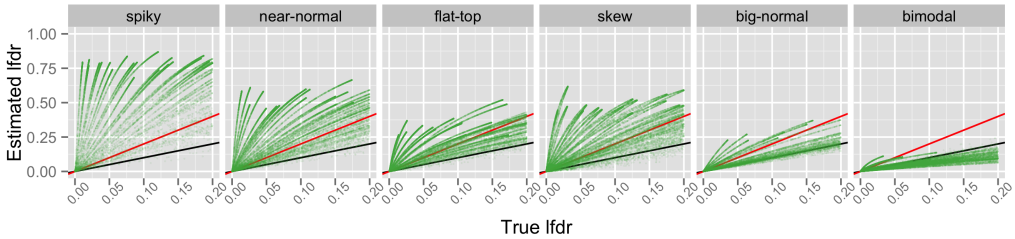
Figure 1. Comparison of the way that different methods decompose p values (left) and z scores (right) into a null component (dark blue) and an alternative component (light blue). In the z score space the alternative distribution is placed on the bottom to highlight the differences in its shape among methods. The three existing methods (`qvalue`, `locfdr`, `mixfdr`) all effectively make the Zero Assumption, which results in a "hole" in the alternative z score distribution around 0. In contrast the method introduced here (`ashr`) assumes that the effect sizes (and thus the z scores) have a unimodal distribution about 0, resulting in a very different decomposition. (In this case the `ashr` decomposition is closer to the truth: the data were simulated under a model where all of the effects are non-zero – specifically, $\beta_j \sim N(0,1)$, $s_j = 1$, resulting in $z_j = \hat{\beta}_j \sim N(0,2)$ – so the "true" decomposition would make everything light blue.)



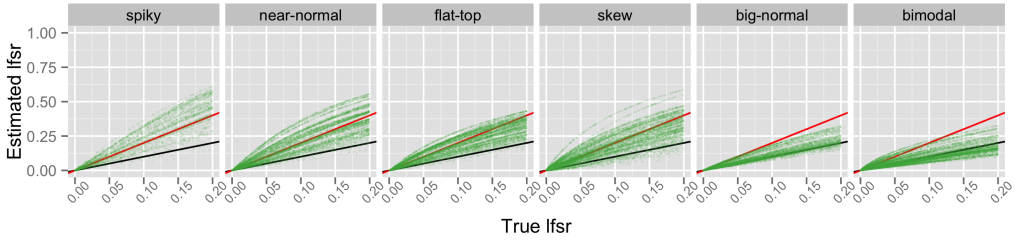
(a) Densities g_1 used in simulations.



(b) Comparison of true and estimated values of π_0. When the UA holds all methods yield conservative (over-)estimates for π_0, with `ashr` being least conservative, and hence most accurate. When the UA does not hold ("bimodal" scenario) the `ashr` estimates are slightly anti-conservative.



(c) Comparison of true and estimated lfdr from `ashr` (ash.n). Black line is $y = x$ and red line is $y = 2x$. Estimates of lfdr are conservative when UA holds, due to conservative estimates of π_0.



(d) As in c), but for lfsr instead of lfdr. Estimates of lfsr are consistently less conservative than lfdr when UA holds, and also less anti-conservative in bimodal scenario.

Figure 2. Results of simulation studies (constant precision $s_j = 1$).

3. For $j = 1, \ldots, J$, simulate $\hat{\beta}_j | \beta_j \sim N(\beta_j, 1)$.

Thus these simulations assume the same precision 1 for each measurement, and for all methods this precision is assumed known (i.e. $\hat{s}_j = 1$).

Figure 2b compares estimates of π_0 from `qvalue`, `locfdr`, `mixfdr` and `ashr` (y axis) with the true values (x axis). For `ashr` we show results for g_1 modelled as a mixture of normal components ("ash.n") and as a mixture of symmetric uniform components ("ash.u"). (Results using the asymmetric uniforms, which we refer to as "half-uniforms", and denote "ash.hu" in subsequent sections, are here generally similar to ash.u and omitted to avoid over-cluttering figures.) The results show that `ashr` provides the smallest more accurate, estimates for π_0, while remaining conservative in all scenarios where the UA holds. When the UA does not hold ("bimodal" scenario) the `ashr` estimates can be slightly anti-conservative. We view this as a minor concern in practice, since we view such a strong bimodal scenario as unlikely in most applications where FDR methods are used. (In addition, the effects on lfsrestimates turn out to be relatively modest; see below).

The lfsr is more robust than lfdr

The results above show that `ashr` can improve on existing methods in producing smaller, more accurate, estimates of π_0 (and, consequently, more accurate estimates of False discovery rates). Nonetheless, in many scenarios `ashr` continues to substantially over-estimate π_0 (see the "spiky" scenario for example). The explanation for this is that these scenarios include an appreciable fraction of "small non-null effects" that are essentially indistinguishable from 0, making accurate estimation of π_0 impossible. Put another way, and as is well known, π_0 is not identifiable: the data can effectively provide an upper bound on plausible values of π_0, but not a lower bound (because the data cannot rule out that everything is non-null, but with miniscule effects). If we desire conservative behaviour we have to estimate π_0 by its upper bound, which can be substantially larger than the true value. Since FDR-related quantities depend quite sensitively on π_0, the consequence of this overestimation of π_0 is corresponding overestimation of FDR (and lfdr, and q values). To illustrate, Figure 2c compares the estimated lfdr from ash.n with the true value (computed using Bayes rule from the true g_1 and $\pi 0$). As predicted, lfdr is overestimated, especially in scenarios where π_0 is overestimated. All methods are similarly affected by this, and the ones that most grossly overestimate π_0 also most grossly overestimated lfdrand FDR/q-values (not shown).

The key point we want to make here is estimation of π_0, and the accompanying identifiability issues, become substantially less troublesome if we use the local false sign rate lfsr(Section **??**), rather than lfdr, to measure significance. This is essentially because lfsris less sensitive to the estimate of π_0. To illustrate, Figure 2 compares the estimated lfsr from ash.n with the true value: although the estimated lfsrcontinue to be conservative, overestimating the truth, the overestimation is substantially less pronounced than for the lfdr, especially for the "spiky" scenario. Further, in the bi-modal scenario, the anti-conservative behaviour is less pronounced in lfsrthan lfdr.

We emphasise that, compared with previous debates regarding estimation vs testing, this section advances an additional reason for favoring focussing on the sign of an effect rather than testing whether it is 0. As noted above, many authors have argued to focus on estimation

because it is implausible that effects are *exactly* 0. Here we add that *even if one believes that some effects may be exactly zero*, it is still better to focus on the sign, because generally *the data are more informative about that question* and so inferences are more robust to, say, the inevitable mis-estimation of π_0. To provide some intuition, consider an observation with a z score of 0. The lfdrof this observation can range from 0 (if $\pi_0 = 0$) to 1 (if $\pi_0 = 1$). But, assuming a symmetric g, the lfsr > 0.5 whatever the value of π_0, because the observation $z = 0$ says nothing about the sign of the effect. Thus, are two reasons to use the lfsr instead of the lfdr: it answers a question that is more generally meaningful (e.g. it applies whether or not zero effects truly exist), and estimation of lfsr is more robust.

Finally for this section, we briefly consider the question of whether we could obtain better estimates of lfsr than are provided here. We speculate that to do this we would either have to i) relax our relatively stringent requirement for conservative behaviour, or ii) be willing to make the assumption that no effects are exactly 0. In support of i), note that under all scenarios the cloud of estimates for lfsr roughly stop at the line $y = x$, suggesting that a reduction in stringency (e.g. reducing the penalty term 15) would risk crossing this line more frequently.

discussion on whether we really need a point mass.

Numerical Stability

The EM algorithm, which we use here to fit our model, is notorious for convergence to local optima. However, in this case, over hundreds of applications of the procedure, we observed no obvious serious problems caused by such behaviour. To quantify this, we ran `ashr` 10 times on each of the 600 simulated datasets above using a random initialization for π, in addition running it using our default initialization procedure (see Implementation details). We then compared the largest log-likelihood achieved across all 11 runs with the log-likelihood achieved by the default run. When using a mixture of normals (ash.n) the results were extremely stable: 96% showed a negligible log-likelihood difference (< 0.02), and the largest difference was 0.8. When using mixtures of uniforms (ash.u,ash.hu) results were slightly less stable: 89% showed a negligible log-likelihood difference (< 0.02), and 6% of runs showed an appreciable log-likelihood difference (> 1), with the largest difference being 5.0. However, perhaps suprisingly, even for this largest difference results from the default run (e.g. the lfsr values, and the posterior means) were in other ways virtually indistinguishable from the results from the run with the highest log-likelihood `github.com/stephens999/ash/dsc-robust/summarize_dsc_robust.rmd`.

The UA helps provide reliable estimates of g

Although we have focussed on estimating the FDR, an important advantage of our EB approach based on modelling the effects β_j (rather than p values or z scores) is that it can estimate the *size* of each effect β_j. Specifically, it provides a posterior distribution for each β_j, which can be used to construct interval estimates for β_j and address question such as "which effects exceed T", for any threshold T. Further, because the posterior distribution is, by definition, conditional on the observed data, interval estimates based on posterior distributions are also valid Bayesian inferences for any subset of the effects that have been selected based on the observed data. This kind of "post-selection" validity is much harder to achieve in the frequentist paradigm.

In particular the posterior distribution solves the (Bayesian analogue of the) "False Coverage Rate" problem posed by [12] which [6] summarizes as follows: "having applied FDR methods to select a set of nonnull cases, how can confidence intervals be assigned to the true effect size for each selected case?". [6] notes the potential for EB approaches to tackle this problem, and [13] consider in detail the case where the non-null effects are normally distributed.

The ability of the EB approach to provide valid "post-selection" interval estimates is extremely attractive in principle. But its usefulness in practice depends on reliably estimating the distribution g. Estimating g is a "deconvolution problem", which are notoriously difficult to solve in general. Indeed, Efron emphasises the difficulties of implementing a stable general algorithm, noting in his rejoinder "the effort foundered on practical difficulties involving the perils of deconvolution... Maybe I am trying to be overly nonparametric ... but it is hard to imagine a generally satisfactory parametric formulation..." ([6] rejoinder, p46). Our key point here is that the UA greatly simplifies the deconvolution problem. While not meeting Efron's desire for an entirely general nonparametric approach, we believe that the UA can handle many cases of practical interest.

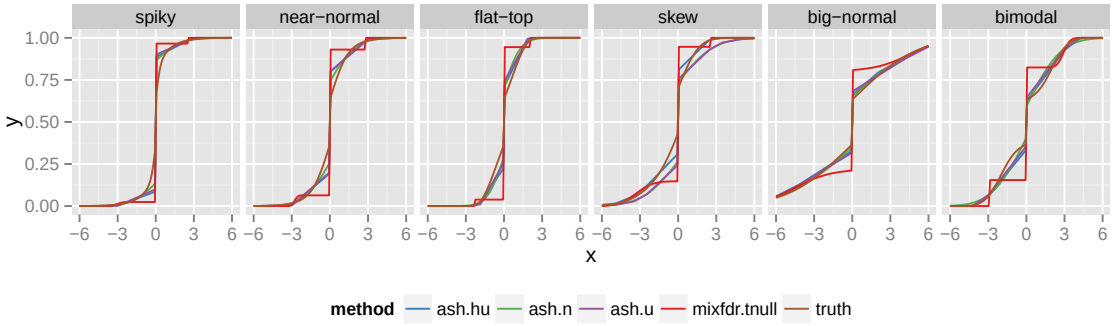
To illustrate this, Figure 3 compares the estimated g from `ashr` with that from `mixfdr` which does not make the UA (and which models g as a mixture of J normal distributions, with $J = 3$ by default). The greater reliability of estimates afforded by the UA is immediately apparent. In particular the estimated cdf from `mixfdr` often has an almost-vertical segment at some non-zero location, indicative of a concentration of density in the estimated g at that location. The UA prevents this kind of "irregular" behaviour, effectively requiring g to be somewhat smooth. While the UA is not the only way to achieve this, we find it an attractive, simple and effective approach.

Interestingly, even in the "bimodal" scenario `ashr` is visually more accurate than `mixfdr`: although `mixfdr` is capable, in principle, of fitting the multiple modes of g, it does not do this well here. Possibly the noise level here is sufficiently large to make reliable estimation of the multiple modes difficult. Indeed, in multi-modal simulations where the multiple modes are sufficiently well-spaced to be clearly visible in the observed $\hat{\beta}$, `mixfdr` fits these modes (Supplementary Information; `dsc-shrink/check_mixfdr_lownoise.rmd`). Of course, we would not advocate the UA in settings where multi-modality is clearly visible in the observed $\hat{\beta}$.

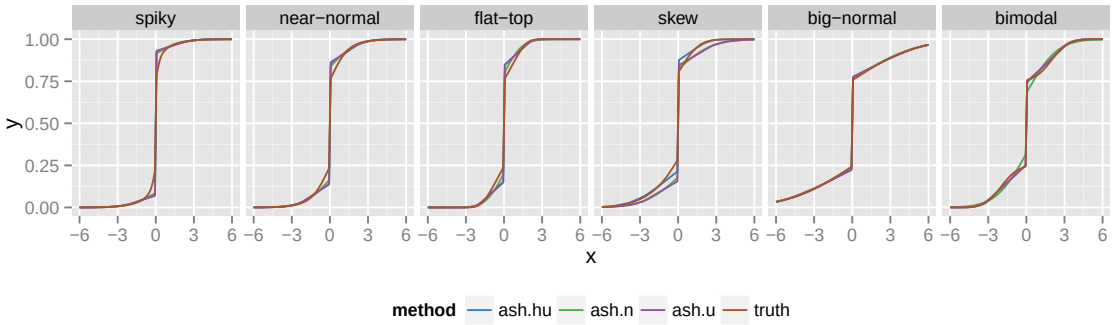
We note one caveat on the accuracy of estimated g: due to the penalty term (15) `ashr` tends to systematically overestimate the mass of g near zero. On careful inspection, this is apparent in Figure 3: the estimated cdf is generally below the true cdf just to the left of zero, and above the true cdf just to the right of zero. Averaging the cdf over many replicates confirms this systematic effect (Figure 3b), and applying our methods without the penalty term removes this systematic effect, although at the cost of sometimes under-estimating π_0 (Figure 3c).

Calibration of posterior intervals

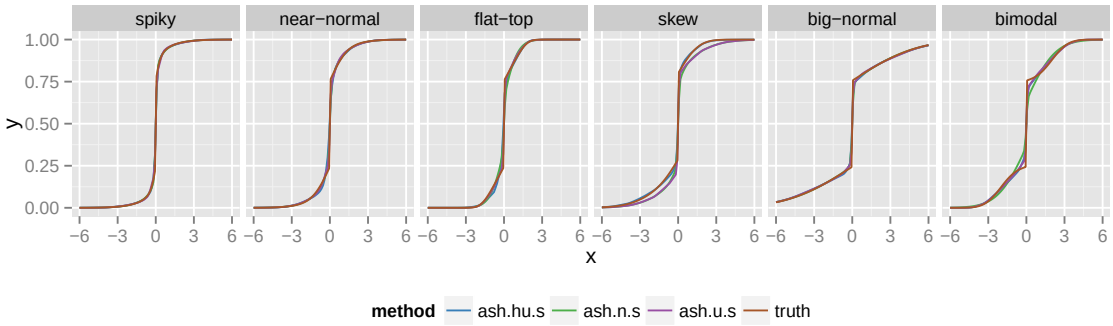
To quantify the effects of errors in estimates of g we examine the calibration of the resulting posterior distributions (averaged over 100 simulations in each Scenario). Specifically we examine the empirical coverage of nominal lower 95% credible bounds for a) all observations; b) significant negative discoveries; c) significant positive discoveries. We examine only lower bounds because the results for upper bounds follow by symmetry (except for the one asymmetric scenario). We



(a) Example estimated cdfs for single data sets compared with truth. The unimodal assumption made by the ash methods effectively regularizes estimates compared with `mixfdr`.



(b) Average estimated cdfs across ~ 10 data sets compared with truth; methods here use penalty (15) so π_0 is systematically overestimated.



(c) Average estimated cdfs across ~ 10 data sets compared with truth; methods here do not use penalty (15) so π_0 is not systematically overestimated. Modest systematic differences from the truth in "skew" and "bimodal" scenarios highlight the modest effects of model mis-specification.

Figure 3. Comparisons of estimated cdfs of g and true cdf of g. See Figure 2b for simulation scenarios.

separately examine positive and negative discoveries because the lower bound plays a different role in each case: for negative discoveries the lower bound is typically large and negative and limits how big (in absolute value) the effect could be; for positive discoveries the lower bound is positive, and limits how small (in absolute value) the effect could be. Intuitively, the lower bound for negative discoveries depends on the accuracy of g in its tail, whereas for positive discoveries it is more dependent on the accuracy of g in the center.

The results are shown in Table 1. Most of the empirical coverage rates are in the range 0.92-0.96 for nominal coverage of 0.95, which we view as adequate for practical applications. The strongest deviations from nominal rates are noted and discussed in the table captions.

	spiky	near-normal	flat-top	skew	big-normal	bimodal
ash.n	0.90	0.94	0.95	0.94	0.96	0.96
ash.u	0.87	0.93	0.94	0.93	0.96	0.96
ash.hu	0.88	0.93	0.94	0.94	0.96	0.96

(a) All observations. Coverage rates are generally satisfactory, except for the extreme "spiky" scenario. This is due to the penalty term (15) which tends to cause over-shrinking towards zero. Removing this penalty term produces coverage rates closer to the nominal levels for uniform and normal methods (not shown). Removing the penalty in the half-uniform case is not recommended (see Appendix).

	spiky	near-normal	flat-top	skew	big-normal	bimodal
ash.n	0.93	0.94	1.00	0.94	0.95	0.98
ash.u	0.86	0.88	0.93	0.91	0.94	0.94
ash.hu	0.87	0.87	0.92	0.93	0.94	0.94

(b) "Significant" negative discoveries. Coverage rates are generally satisfactory, except for the uniform-based methods in the spiky and near-normal scenarios, and the normal-based method in the flat-top scenario. These results likely reflect inaccurate estimates of the tails of g due to a disconnect between the tail of g and the component distributions in these cases. For example, the uniform methods sometimes substantially underestimate the length of the tail of g in these long-tailed scenarios, causing over-shrinkage of the tail toward 0.

	spiky	near-normal	flat-top	skew	big-normal	bimodal
ash.n	0.94	0.94	0.94	0.86	0.95	0.96
ash.u	0.93	0.93	0.93	0.84	0.95	0.95
ash.hu	0.92	0.92	0.93	0.92	0.95	0.95

(c) "Significant" positive discoveries. Coverage rates are generally satisfactory, except for the symmetric methods under the asymmetric ("skew") scenario.

Table 1. Table of empirical coverage for nominal 95% lower credible bounds

Differing measurement precision across units

We turn now to the second important component of our work: allowing for varying measurement precision across units. The key to this is the use of a likelihood, (4) or (12), that explicitly

incorporates the measurement precision (standard error) of each $\hat{\beta}_j$.

To illustrate, we conduct a simulation where half the measurements are quite precise (standard error $s_j = 1$), and the other half are very poor ($s_j = 10$). In both cases, we assume that half the effects are null and the other half are normally distributed with standard deviation 1:

$$p(\beta) = 0.5\delta_0(\beta) + 0.5N(\beta; 0, 1). \tag{14}$$

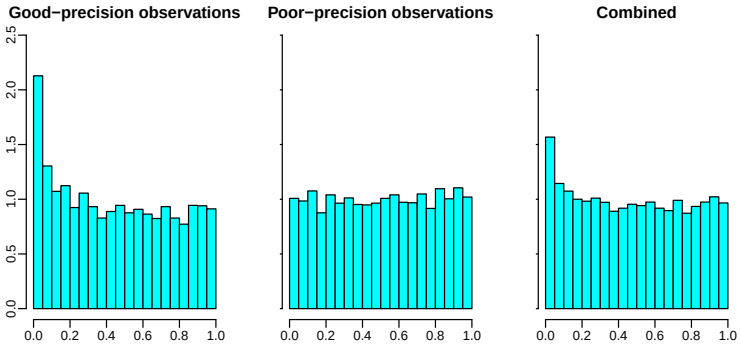
In this setting, the poor-precision measurements tell us very little, and any sane analysis should effectively ignore them. However, this is not the case in standard FDR-type analyses (Figure 4). This is because the poor measurements produce p values that are approximately uniform (Figure 4a), which, when combined with the good-precision measurements, dilute the overall signal (e.g. they reduce the density of p values near 0). This is reflected in the results of FDR methods like `qvalue` and `locfdr`: the estimated error rates (q-values, or lfdr values) for the good-precision observations increase when the low-precision observations are included in the analysis (Figure 4b). In contrast, the results from `ashr` for the good-precision observations are unaffected by including the low-precision observations in the analysis (Figure 4b).

Reordering of significance, and the "p value prior"

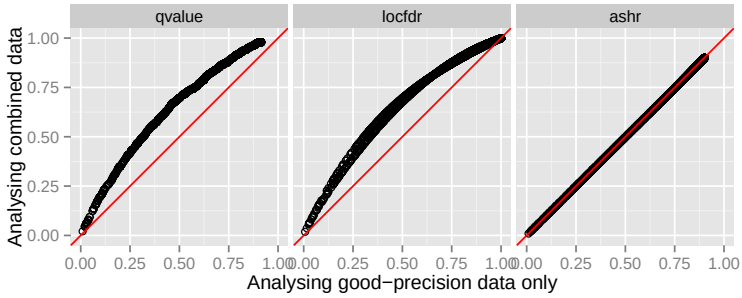
Another important consequence of accounting for varying measurement precision across units is that `ashr` re-orders the significance of the observations compared with the original p values or z scores. This is illustrated, using the same simulation as above, in Figure 5 (left panel). We see that poor precision measurements are assigned a higher lfsr than good precision measuments that have the same p value. The intuition is that, due to their poor precision, these measurements contain very little information about the sign of the effects (or indeed any other aspect of the effects), and so the lfsr for these poor-precision measurements is always high.

This tendency to reorder the significance of observations, and specifically to downweight the significance of observations with low precision, is a consequence of an assumption that we have made up to now: that the distribution of the effect β_j is independent of its standard error s_j (equation (**??**)). One way to relax this assumption is to allow β_j to scale with s_j^α for some power α; that is, to assume model (11). Setting $\alpha = 0$ implies that β_j is independent of s_j, as we have assumed up to now, and $\alpha > 0$ implies that observations with larger standard error tend to have larger effects (in absolute value). This latter assumption may often be qualitatively plausible: for example, in gene expression studies the standard error for gene j depends partly on the variance of its expression among samples, and genes with a larger variance may tend to be less tightly regulated and so be amenable to a larger shift in expression between conditions (i.e. larger effect β_j). The special case $\alpha = 1$ corresponds to assuming that the z scores $z_j = \beta_j/s_j$, and hence the corresponding p values, are identically distributed, independent of s_j. This is essentially the assumption made (implicitly or explicitly) by existing methods like `locfdr`, `mixfdr` and `qvalue`, which model the z_j or p_j directly. Under this assumption, and with the normal mixture prior for β_j/s_j (3), the lfsrcomputed by `ashr` provides the same *ranking* of observations as the z scores and p values (see Figure 5, right panel).

This observation extends a result in [21], who showed that, when testing a series of null hypotheses, with a normal likelihood, Bayes Factors and p values produce the same ranking of tests *if* the Bayes Factor for the jth test is computed using a normal prior for the alternative



(a) Density histograms of p values for good-precision, poor-precision, and combined observations



(b) Comparison of results of different methods applied to good-precision observations only (x axis) and combined data (y axis). Each point shows the "significance" (q values from `qvalue`; lfdr for `locfdr`; lfsr for `ashr`) of a good-precision observation under the two different analyses.

Figure 4. Simulation illustrating how, for existing FDR methods, poor-precision observations can contaminate signal from good-precision observations. The top panel (a) illustrates that when p values from good-precision observations (left) and from poor-precision observations (center) are combined (right), they produce a distribution of p values with less overall signal - and so, by conventional methods, will give a higher estimated FDR at any given threshold. The bottom panel (b) illustrates this behaviour directly for the methods `qvalue` and `locfdr`: the q-values from `qvalue` and the lfdr estimates from `locfdr` are higher when applied to all data than when applied to good-precision observations only. In contrast the methods described here (`ashr`) produce effectively the same results (here, the lfsr) in the good-precision and combined data analyses.

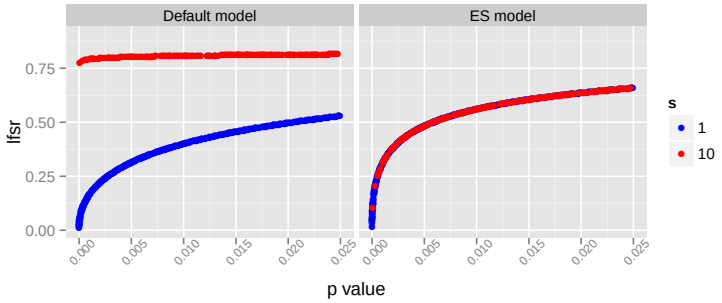


Figure 5

effects with mean 0 and variance Ks_j^2 for some constant K. Because of this correspondance in ranking, Wakefield refers to this as the "p-value prior". Our result extends Wakefield's results to the mixture of normals prior, from a single normal. As an aside we note that the result does depend on the normal assumptions: for example, it does not hold in general for mixtures of uniforms as a prior; nor if we use a t likelihood instead of a normal likelihood. That is, under these models, the EB analysis may reorder the significance of the observations (compared with the p values) even when $\alpha = 1$.

To summarize: the behaviour of `ashr` depends on the value of α assumed in (11). Up to now we have assumed $\alpha = 0$, which implies that effects are identically distributed, independent of their standard error. In contrast, existing methods effectively assume $\alpha = 1$, which implies that effects tend to scale proportional to their standard error. Whether it is better to set $\alpha = 0$ or $\alpha = 1$ in practice will depend on actual relationship between β_j and s_j, which will be dataset-specific. Indeed, it seems quite likely that, in general, the optimal α may be some other (non-integer) value. Framing the problem in this way – that is, as comparing different modelling assumptions for β_j, rather than as comparing "modelling β_j" vs "modelling z_j" (or "modelling p_j") – has the important advantage that likelihood-based methods can be used to select α. For example, following the logic of the Empirical Bayes approach it would be natural to select α by maximum likelihood. Since α is a one-dimensional parameter, this can be achieved by a 1-d grid search, which has been implemented in our software by C. Dai. (Alternatively, since fractional values of α may be tricky to interpret, users may prefer to simply choose whichever of $\alpha = 0, 1$ yields the highest likelihood.)

Adaptive Shrinkage

Johnstone and Silverman [22] (see also [33]) note the ability of EB methods to adapt to overall signal strength when attempting to identify and estimate strong signals. The idea is that, if most β_j are truly at or near zero then (given enough data) the estimated prior distribution $\hat{g}$ will reflect this by being concentrated near 0, and the posterior distributions $p(\beta_j|\hat{\beta}_j, \hat{s}_j, \hat{g})$ will consequently also tend to concentrate around zero (strong shrinkage). In contrast, if most β_j are large then $\hat{g}$ will be flatter, resulting in less shrinkage. Indeed, in the limit as $\hat{g}$ becomes flatter and flatter the posterior distribution on β_j becomes $N(\hat{\beta}_j, \hat{s}_j)$, and the Bayesian credible intervals match standard confidence intervals, which might be considered "no shrinkage".

Here we build on this idea in two ways. First, by using flexible semi-parametric approaches to estimate g we aim to maximise the potential for this adaptive behaviour. Second, by incorporating the precision of each measurement into the likelihood, (4) or (12), we ensure that shrinkage adapts to measurement precision: more precise measurements undergo less shrinkage than less precise measurements. This is because observations $\hat{\beta}_j$ with larger standard errors have flatter likelihoods than observations with small standard error, and so their posteriors will be more affected by the prior, which, being unimodal at 0, tends to shrink estimates towards 0. To emphasise these two key features we refer to our method as "**a**daptive **sh**rinkage". Figure 6 illustrates the idea by contrasting results for two simulation scenarios with different signal strengths, where observations vary in their measurment precision.

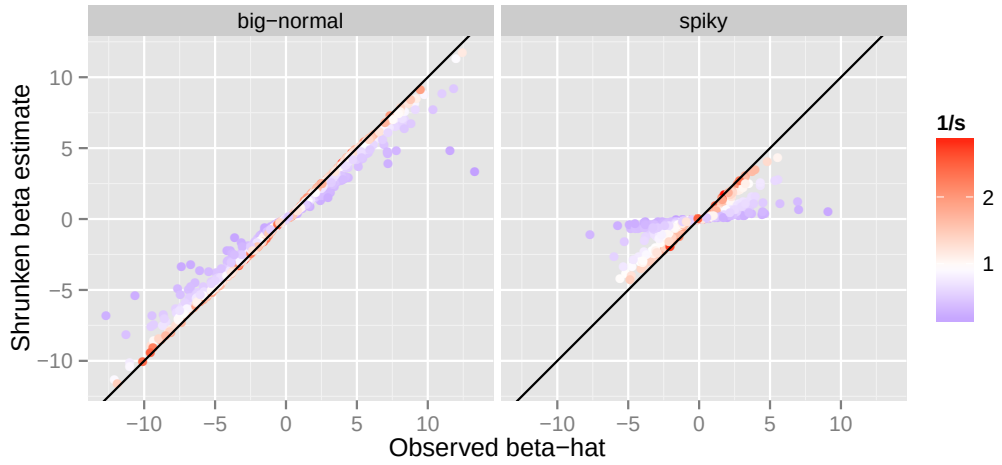


Figure 6. Figure illustrating adaptive shrinkage. Results are shown for two different scenarios: (left) "big-normal" where the effects have a wide distribution; (right) "spiky", where the effects are more concentrated near 0. The standard error of each observation, s_j, was simulated from Inverse Gamma(5,5), which resulted in s_j varying from 0.34 to 6.7. The shrunken estimates (y axis) are plotted against the observed estimates (x axis), with color indicating the (square root of the) precision of each measurement (blue=lower-precision; red=higher-precision). The two key features are i) shrinkage is adaptive to signal in data, so shrinkage is stronger for the spiky scenario; ii) shrinkage is adaptive to the precision of each measurement, so less precise observations (blue) are shrunken more strongly.

0.1 Discussion

The use of EB methods for estimation also has a long and close relationship with another key concept in modern statistics: shrinkage estimation [15]. This close relationship becomes especially intimate here due to our unimodal assumption on the underlying effects, which encourages shrinkage towards the mode of the distribution. Thus, although we focus here primarily on FDR-related issues, another important contribution of our work is to provide *generic* and *adaptive* shrinkage estimation procedures. By generic, we mean that these methods can be applied in any setting where a series of effect estimates and corresponding standard errors are available. By adaptive, we have two properties in mind. First, the appropriate amount of shrinkage is determined from the data: when the data indicate that large effects are relatively common then shrinkage of large observed effects is less than when the data indicate that large effects are rare. Second, the amount of shrinkage undergone by each measurement will depend on its precision, or standard error: measurements with big standard errors undergo more shrinkage than measurements with small standard errors. Shrinkage estimation is a powerful tool, with many potential applications, including for example wavelet denoising, where shrinkage of wavelet coefficients is used to smooth signals. Our methods provide an attractive and competitive alternative to existing EB shrinkage methods for that context, as will be explored more fully elsewhere (Xing and Stephens, in preparation).

Choice of mixture components. Use of likelihood to compare different mixture components?

An important focus of Efron's work, also implemented in `locfdr`, is to relax the assumption that the null z scores are $N(0,1)$, to allow for empirical deviations from this theoretical null distribution. Although important, this issue is largely orthogonal to the issues we focus on here, which arise even if the theoretical null distribution holds precisely in the empirical data, and so we consider it just briefly in the Discussion.

noting that the p values in Figure **??** were actually generated under a scenario where *no tests were null*! Thus the qvalue procedure, which estimates the proportion of nulls to be xx%, is acting very conservatively in this case. However, it could be argued that some degree of conservativeness is inevitable, and that nonetheless the procedure is acting sensibly. Therefore, to illustrate this point we really want to focus on the fact that, at least in some settings, the assumption is *unrealistic*. Specifically, we want to argue that the implied distribution of the alternative p values is unrealistic. In fact, this is hard to see in p value space, and so instead we translate these p values back to the z scores that lead to these p values.

- it allows us to easily just vary sigma on a grid, and fit pi, which makes allowing different noise levels really easy! In contrast, incorporating differential errors messes up the EM algorithm if we try to estimate σ_k.

Directly modeling the p values, or z scores, say via non-parametric methods, can lead to unrealistic distributions being fitted. Put another way, because z scores are the result of adding noise to some distribution, the range of distributions they can take is limited Using entirely non-parametric methods loses this information. The solution is to model beta hat as a convolution of some distribution g and an error component.

About benefits of generic approach. In outline, our goal is to perform Bayesian inference for the true effects β using information in the summary statistics $\hat{\beta}, \hat{s}$. That is we aim to compute the joint posterior distribution $p(\beta|\hat{\beta}, \hat{s})$. This is connected with work by [21] (see also [34]) who

performs Bayesian inference from summary statistics, computing $p(\beta_j|\hat{\beta}_j, \hat{s}_j)$ for a single j, with a specific fixed prior on β_j. Essentially we extend this to multiple j, using a hierarchical model to connect the observations. By working with summary statistics $\hat{\beta}, \hat{s}$, rather than more refined data, we aim to produce generic methods that can be applied whenever such summary data are available - just as `qvalue` can be applied to any set of p values for example. Any attempt to produce generic methods is likely to involve a compromise between functionality and generality; we believe that by working with two numbers $(\hat{\beta}_j, \hat{s}_j)$ for each observation, rather than one (p_j or z_j), we can gain substantially in functionality (e.g. we can estimate effect sizes, as well as testing, and we can take better account of variations in measurement precision across units j) while losing only a little in generality.

Before proceeding, we feel obliged to correct the common misconception that the use of FDR methods is to "correct" for the number of tests performed. In fact the FDR does not depend on the number of tests. This is because, as the number of tests increases, both the true positives and false positives increase linearly, and the FDR remains the same. (If this intuitve argument does not convince, see [20], and note that the FDR at a given p value threshold does not depend on the number of tests m.) A better way to think of it is that the FDR accounts for the *amount of signal* in the tests that were performed: if there are lots of strong signals then the FDR at a given threshold may be low, even if a large number of tests were performed; and conversely if there are no strong signals then the FDR at the same threshold may be high, even if relatively few tests were performed. In other words, FDR corrects for the *results* of all tests performed, not the *number* of all tests performed.

Note on multiple comparisons: it isn't really a "problem" but an "opportunity". This viewpoint also espoused by [2]. It isn't the number of tests that is relevant (the false discovery rate at a given threshold does not depend on the number of tests). It is the *results* of the tests that are relevant. Performing multiple comparisons, or multiple tests, is often regarded as a "problem". However, here we regard it instead as an opportunity - an opportunity to combine (or "pool") information across tests or comparisons. Focussing on the number of tests performed can be seen as an approximation.

Here performing inference means obtaining an (approximate) posterior distribution for each effect β_j, which can be used to estimate False Discovery Rates, and to produce both point and interval estimates for β_j. Although "multiple comparisons" is usually presented as a "problem", our emphasis here is that the multiple measurements actually provide an opportunity: the hierarchical model combines information across measurements to improve both accuracy and precision of estimates.

There are two issues with existing approaches to FDR estimation and control that I would like to address here. The first is that they may do not take proper account of differences in measurement precision across units (i.e. differences in $\hat{s}_j$). The second is that the Zero Assumption, although initially appealing as a "conservative" assumption, is often an unrealistic and unnecessarily conservative assumption. As we shall see, both factors can cause the FDR to be overestimated.

[6] states the Zero Assumption as the assumption that "most of the z-values near zero come from null genes". His main aim in making this assumption is to estimate an empirical null though (not assume N(0,1) for the null) rather than to impose identifiability.

Note that [11] models z scores as something plus noise under both H_0 and H_1, which avoids

this problem. (DOes the same maybe apply to modeling beta, rather than z scores, when the errors vary?)

Rice and Spiegelhalter - BRCA data?

A fundamental idea is that the measurements of β_j for each gene can be used to improve inference for the values of β for other genes.

— NOte In principle we might prefer a full Bayes approach that accounts for uncertainty in π (or, more generally, in g); however, we believe that in most practical applications uncertainty in g will not be the most important concern, and compromise this principle for the simplicity of the EB approach.

Implementation details

Choice of grid for σ_k, a_k

When f_k is $N(0, \sigma_k)$ we specify our grid by specifying: i) a maximum and minimum value $(\sigma_{\min}, \sigma_{\max})$; ii) a multiplicative factor m to be used in going from one gridpoint to the other, so that $\sigma_k = m\sigma_{k-1}$. The multiplicative factor affects the density of the grid; we used $m = \sqrt{2}$ as a default. We chose $\sigma_{\min}$ to be small compared with the measurement precision ($\sigma_{\min} = min(\hat{s}_j)/10$) and $\sigma_{\max} = 2\sqrt{\max(\hat{\beta}_j^2 - \hat{s}_j^2)}$ based on the idea that $\sigma_{\max}$ should be big enough so that $\sigma_{\max}^2 + \hat{s}_j^2$ should exceed $\hat{\beta}_j^2$. (In rare cases where $\max(\hat{\beta}_j^2 - \hat{s}_j^2)$ is negative we set $\sigma_{\max} = 8\sigma_{\min}$.)

When the mixture components f_k are uniform, we use the same grid for the parameters a_k as for σ_k described above.

Our goal in specifying a grid was to make the limits sufficiently large and small, and the grid sufficiently dense, that results would not change appreciably with a larger or denser grid. For a specific data set one can of course check this by experimenting with the grid, but these defaults usually work well in our experience.

Penalty term on π

In practice data do not distinguish between effects that are "very small" and "exactly zero". Thus, if a mixture component f_k has most mass very near zero then this introduces non-identifiability between π_k and π_0. In practice this is a major problem only if one focuses on false discovery rates rather than false sign rates: the lfdris sensitive to π_0 whereas the lfsris not. However, to make lfdr(and lfsr) estimates from our method "conservative" we add a penalty term $h(\pi; \lambda)$ to the likelihood to encourage over-estimation of π_0:

$$h(\pi; \lambda) = \prod_{k=0}^{K} \pi_k^{\lambda_k - 1} \tag{15}$$

where $\lambda_k \geq 1 \, \forall k$. The default is $\lambda_0 = 10$ and $\lambda_k = 1$, which yielded consistently conservative estimation of π_0 in our simulations (Figure **??**).

Although this penalty is based on a Dirichlet density, we do not interpret this as a "prior distribution" for π: we chose it to provide conservative estimates of π_0 rather than to represent prior belief.

Likelihood for π and EM algorithm

To formally derive the likelihood we restate our modelling assumptions more formally. We provide details first for the normal likelihood and then briefly describe modifications for the t likelihood.

We treat the standard errors as $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_J)$ as given, and perform all inference conditional on $\hat{s}$. We assume that given $\hat{s}$ and the hyper-parameters π, the $(\beta_j, \hat{\beta}_j)$ pairs are independent

$$p(\beta_j \mid \hat{s}, \pi) = \sum_{k=0}^{K} \pi_k f_k(\beta_j) \tag{16}$$

$$p(\hat{\beta}_j \mid \beta_j, \hat{s}, \pi) = N(\hat{\beta}_j; \beta_j, \hat{s}_j^2), \tag{17}$$

where for notational convenience we have replaced the point mass δ_0 with f_0.

Multiplying (16) and (17) together gives $p(\hat{\beta}_j, \beta_j | \hat{s}, \pi)$, which integrating over β_j yields

$$p(\hat{\beta}_j | \hat{s}, \pi) = \sum_{k=0}^{K} \pi_k \tilde{f}_k(\hat{\beta}_j) \tag{18}$$

where

$$\tilde{f}_k(\hat{\beta}_j) := \int f_k(\beta_j) N(\hat{\beta}_j; \beta_j, \hat{s}_j^2)\, d\beta_j \tag{19}$$

denotes the convolution of f_k with the normal density. These convolutions are straightforward to evaluate whether f_k is a normal or uniform density. Specifically,

$$\tilde{f}_k(\hat{\beta}_j) = \begin{cases} N(\hat{\beta}_j; 0, \hat{s}_j^2 + \sigma_k^2) & \text{if } f_k(\cdot) = N(\cdot; 0, \sigma_k^2), \\ \frac{\Psi((\hat{\beta}_j - a_k)/\hat{s}_j) - \Psi((\hat{\beta}_j - b_k)/\hat{s}_j)}{b_k - a_k} & \text{if } f_k(\cdot) = U(\cdot; a_k, b_k), \end{cases} \tag{20}$$

where Ψ denotes the distribution function of the standard normal distribution. If we use a t_ν likelihood instead of a normal likelihood then the convolution is tricky for f_k normal and we have not implemented it; for f_k uniform the result (**??**) holds but with the standard normal distribution function replaced with the t_ν distribution function.

With this in place, the likelihood for π is obtained by multiplying across j:

$$L(\pi) = \prod_j \sum_k \pi_k w_{kj} \tag{21}$$

where the $l_{kj} := \tilde{f}_k(\hat{\beta}_j)$ are known. We add to this likelihood a penalty term $h(\pi; \lambda)$ above, and use an EM algorithm to maximize $L(\pi) + h(\pi; \lambda)$. The one-step updates for this EM algorithm

are:

$$w_{kj} = \pi_k l_{kj} / \sum_{k'} \pi_{k'} l_{k'j} \tag{22}$$

$$n_k = \sum_j w_{kj} + \lambda_k - 1 \quad \text{E Step} \pi_k = n_k / \sum_{k'} n_{k'} \quad \text{M step} \tag{23}$$

Note that π_k can be interpreted as the prior probability that β_j arose from component k, and l_{kj} is the likelihood for β_j given that it arose from component k, so w_{kj} is the posterior probability that β_j arose from component k, given $\hat{\beta}, \hat{s}, \pi$. Thus n_k is the expected number of β_j that arose from component k, plus pseudo-counts $\lambda_k - 1$ from the penalty term. We used the elegant R package `SQUAREM` [35]) to accelerate convergence of this EM algorithm.

Initialization

By default we initialize our EM algorithm with $\pi_k = 1/n$ for $k = 1, \ldots, K$, with $\pi_0 = 1 - \pi_1 - \cdots - \pi_K$. (In all our simulations here $K << n$ so this initializes with most mass on π_0.) Our rationale to initializing "near the null" like this is that we expect that strong signal in the data can quickly draw the EM algorithm away from the null (in a single iteration), but weak signal in the data cannot quickly draw the algorithm towards the null.

In addition, once the EM algorithm has converged, we check that the (penalized) log-likelihood attained is higher than that achieved by the global null solution $\pi_0 = 1$. If not then we replace the EM solution with the global null solution. (This helps guard against errors due to convergence to a local optimum when the data are consistent with the global null.)

Conditional distributions

Given $\hat{\pi}$, we compute the conditional distributions

$$p(\beta_j | \hat{\pi}, \hat{\beta}, s) \propto g(\beta_j; \pi) L(\beta_j; \hat{\beta}_j, \hat{s}_j). \tag{24}$$

Each posterior is a mixture on $K + 1$ components:

$$p(\beta_j | \hat{\pi}, \hat{\beta}, s) = \sum_{k=0}^{K} w_{kj} p_k(\beta_j | \hat{\beta}_j, \hat{s}_j) \tag{25}$$

where the posterior weights w_{kj} are computed as in (22) with $\pi = \hat{\pi}$, and the posterior mixture component p_k is the posterior on β_j that would be obtained using prior $f_k(\beta_j)$ and likelihood $L(\beta_j; \hat{\beta}_j, \hat{s}_j)$. All these posterior distributions are easily available. For example, if f_k is uniform and L is t_ν then this is a truncated t distribution. If f_k is normal and L is normal, then this is a normal distribution.

Scenario	Alternative distribution, g_1
spiky	$0.4N(0, 0.25^2) + 0.2N(0, 0.5^2) + 0.2N(0, 1^2), 0.2N(0, 2^2)$
near normal	$2/3N(0, 1^2) + 1/3N(0, 2^2)$
flattop	$(1/7)[N(-1.5, .5^2) + N(-1, .5^2) + N(-.5, .5^2) +$ $N(0, .5^2) + N(0.5, .5^2) + N(1.0, .5^2) + N(1.5, .5^2)]$
skew	$(1/4)N(-2, 2^2) + (1/4)N(-1, 1.5^2) + (1/3)N(0, 1^2) + (1/6)N(1, 1^2)$
big-normal	$N(0, 4^2)$
bimodal	$0.5N(-2, 1^2) + 0.5N(2, 1^2)$

Table 2. Summary of simulation scenarios considered

A Definitions of FDR-related concepts used here

Let Γ denote a subset of the tests $1, \ldots, J$ declared "significant" by some procedure. We define the FDR for Γ as the proportion of these tests that are "false discoveries" in that H_j is true:

$$\mathrm{FDR}(\Gamma; H) := \frac{\#\{j : H_j \cap j \in \Gamma\}}{\#\{j : j \in \Gamma\}}, \tag{26}$$

taken to be 0 if both numerator and denominator are 0. Frequentist approaches to controlling the FDR attempt to control the expectation of $\mathrm{FDR}(\Gamma; H)$ where the expecation is taken over the sampling distribution of Γ (which is a function of the procedure used, considered fixed, and the data D, considered random). Bayesian approaches, in contrast, condition on the observed data, and compute the posterior distribution for the H_j given the data, which induces a posterior distribution on $\mathrm{FDR}(\Gamma)$. This posterior distribution can be used to obtain a point estimate for $\mathrm{FDR}(\Gamma)$, for example using the posterior mean. To be more explicit, Bayesian (and EB) approaches give the posterior probability that H_j holds, which analagous to [6] we refer to as the "local FDR", and denote lfdr_j:

$$\mathrm{lfdr}_j := \Pr(H_j|D). \tag{27}$$

The FDR can be estimated by the posterior mean for $\mathrm{FDR}(\Gamma)$, given by

$$\widehat{\mathrm{FDR}}(\Gamma) := \frac{\sum_{j\in\Gamma} \mathrm{lfdr}_j}{\#\{j : j \in \Gamma\}}. \tag{28}$$

Although our approach here is (Empirical) Bayesian, one nice feature of FDR procedures is that Frequentist and Bayesian procedures are often closely aligned, at least under certain assumptions; see for example [20], particularly his Theorem 1.

The q value for observation j is defined [20] as, roughly, the FDR for the set of observations that are at least as significant as observation j. In a Bayesian analysis it is natural to order the significance of the hypotheses by lfdr_j, so we define

qvalue

To perform this decomposition, `qvalue` makes two key assumptions: i) the null p values have a uniform distribution, and ii) the p values near 1 are all from tests where the null is true.

Under these assumptions performing the decomposition boils down to drawing the horizontal line on the p value histogram that intersects with the distribution at $p = 1$: the area under this horizontal line defines the uniform component that is due to the null p values, with the remainder necessarily corresponding to the alternative p values. In particular, the height of the horizontal line is an estimate of π_0, the proportion of p values that come from the null. Now, for any given threshold, although we do not know *which* p values correspond to null tests and which to alternative tests, the areas of the two components indicate *how many* p values correspond to each, allowing the FDR to be estimated (Figure **??**). Specifically, at threshold γ, and given an estimate $\hat{\pi}_0$ for π_0, a natural non-parametric estimate of the FDR is

$$\hat{\text{FDR}}(\gamma) = J\hat{\pi}_0\gamma/\#\{j : p_j \leq \gamma\}. \tag{29}$$

See [36], equation (8).

Frequentist approaches to FDR often concentrate on the *average* error rate for a subset of observations whose significance exceeds some threshold. Bayesian versions of such quantities can also be computed. For example, the FDR for any subset of effects $\Gamma \subset \{1, \ldots, J\}$ can be estimated by

$$\widehat{\text{FDR}}(\Gamma) := \frac{\sum_{j\in\Gamma} \text{lfdr}_j}{\#\{j : j \in \Gamma\}}. \tag{30}$$

And, analogous to [20], we can define the q value for observation j (q_j) as an estimate of the FDR for all observations that are at least as significant as j:

$$q_j := \widehat{\text{FDR}}(\{k : \text{lfdr}_k \leq \text{lfdr}_j\}). \tag{31}$$

Since the FDR terminology has become so widespread, we here introduce terminology for these ideas that parallels FDR terminology. Recall that, in FDR parlance, a "discovery" is an effect that is declared to be non-zero; and a "false discovery" refers to any such declaration that is made in error. The FDR is the (expected) proportion of false discoveries among all discoveries. Consider now insisting that for every discovery, we *also declare the sign of the effect* (positive or negative). We call such a declaration a "signed discovery"; and correspondingly a "false signed discovery" is any such declaration that is made in error - that is, any signed discovery that is either 0, or whose sign is opposite to that declared. Based on these definitions we can define the "false signed discovery rate" (or "false sign rate", FSR, for short) as the (expected) proportion of false signed discoveries among all signed discoveries. Similarly, we can define the local false sign rate, lfsr_j to be the probability that observation j would be a false signed discovery, were we to declare it a signed discovery. Or, equivalently, but perhaps more simply, the lfsr_j is the probability we would get the sign of β_j wrong if we were to be forced to declare it either positive or negative. Explicitly:

By analogy with (30) we can also define an estimate of the average False Sign Rate for any subset of effects,

$$\widehat{\text{FSR}}(\Gamma) := \frac{\sum_{j\in\Gamma} \text{lfsr}_j}{\#\{j : j \in \Gamma\}}. \tag{32}$$

And, by analogy with (**??**) we can define the "s-value" for each observation

$$s_j := \widehat{\text{FSR}}(\{k : \text{lfsr}_k \leq \text{lfsr}_j\}) \tag{33}$$

being the average error rate for all observations that are at least as significant as j.

Problems with removing the penalty term in the half-uniform case

We note here an unanticipated problem we came across when using no penalty term in the half-uniform case: when the data are nearly null, the estimated g converges, as expected and desired, to a distribution where almost all the mass is near 0, but sometimes all this mass is concentrated almost entirely just to one side (left or right) or 0. This can have a very profound effect on the local false sign rate: for example, if all the mass is just to the right of 0 then all observations will be assigned a very high probability of being positive (but very small), and a (misleading) low local false sign rate. For this reason we do not recommend use of the half-uniform with no penalty.

Acknowledgements

Statistical analyses were conducted in the R programming language [37], Figures produced using the ggplot2 package [38], and text prepared using LaTeX. Development of the methods in this paper was greatly enhanced by the use of the `knitr` package [39] within the RStudio GUI, and git and github. The `ashr` R package is available from `http://github.com/stephens999/ashr` , and includes contributions from Chaixing (Rick) Dai, Mengyin Lu, and Tian Sen.

This work was supported by NIH grant xxx and a grant from the Gordon and Betty Moore Foundation.

References

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

2. Greenland S, Robins JM (1991) Empirical-bayes adjustments for multiple comparisons are sometimes useful. Epidemiology : 244–251.

3. Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical bayes analysis of a microarray experiment. Journal of the American statistical association 96: 1151–1160.

4. Efron B, Tibshirani R (2002) Empirical bayes methods and false discovery rates for microarrays. Genetic epidemiology 23: 70–86.

5. Efron B, et al. (2003) Robbins, empirical bayes and microarrays. The annals of Statistics 31: 366–378.

6. Efron B (2008) Microarrays, empirical bayes and the two-groups model. Statistical Science 23: 1–22.

7. Efron B (2010) Large-scale inference: empirical Bayes methods for estimation, testing, and prediction, volume 1. Cambridge University Press.

8. Kendziorski C, Newton M, Lan H, Gould M (2003) On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. Statistics in medicine 22: 3899–3914.

9. Newton MA, Noueiry A, Sarkar D, Ahlquist P (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics 5: 155–176.

10. Datta S, Datta S (2005) Empirical bayes screening of many p-values with applications to microarray studies. Bioinformatics 21: 1987–1994.

11. Muralidharan O (2010) An empirical bayes mixture method for effect size and false discovery rate estimation. The Annals of Applied Statistics : 422–438.

12. Benjamini Y, Yekutieli D (2005) False discovery rate–adjusted multiple confidence intervals for selected parameters. Journal of the American Statistical Association 100: 71–81.

13. Zhao Z, Gene Hwang J (2012) Empirical bayes false coverage rate controlling confidence intervals. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74: 871–891.

14. Gelman A, Hill J, Yajima M (2012) Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness 5: 189–211.

15. EFRON B, MORRIS C (1973) A Bayesian derivation of the James-Stein estimator. JASA 68: 117.

16. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, series B 39: 1–38.

17. Tukey JW (1991) The philosophy of multiple comparisons. Statistical science : 100–116.

18. Tukey JW (1962) The future of data analysis. The Annals of Mathematical Statistics : 1–67.

19. Gelman A, Tuerlinckx F (2000) Type s error rates for classical and bayesian single and multiple comparison procedures. Computational Statistics 15: 373–390.

20. Storey J (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. The Annals of Statistics 31: 2013–2035.

21. Wakefield J (2009) Bayes factors for genome-wide association studies: comparison with p-values. Genet Epidemiol 33: 79-86.

22. Johnstone IM, Silverman BW (2004) Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. Annals of Statistics : 1594–1649.

23. Carvalho CM, Polson NG, Scott JG (2010) The horseshoe estimator for sparse signals. Biometrika : asq017.

24. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, et al. (2015) Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet 11: e1004969.

25. Cordy CB, Thomas DR (1997) Deconvolution of a distribution function. Journal of the American Statistical Association 92: 1459–1465.

26. Xie X, Kou S, Brown LD (2012) Sure estimates for a heteroscedastic hierarchical model. Journal of the American Statistical Association 107: 1465–1479.

27. Sarkar A, Mallick BK, Staudenmayer J, Pati D, Carroll RJ (2014) Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. J Comput Graph Stat 23: 1101-1125.

28. Koenker R, Mizera I (2014) Convex optimization, shape constraints, compound decisions, and empirical bayes rules. Journal of the American Statistical Association 109: 674–685.

29. Jiang W, Zhang CH (2009) General maximum likelihood empirical bayes estimation of normal means. The Annals of Statistics 37: 1647–1684.

30. Khintchine AY (1938) On unimodal distributions. Izv Nauchno-Isled Inst Mat Mech Tomsk Gos Univ 2: 1–7.

31. Shepp L (1962) Symmetric random walk. Transactions of the American Mathematical Society : 144–153.

32. Feller W (1971) An introduction to probability and its applications, vol. ii. Wiley, New York .

33. Raykar VC, Zhao LH (2011) Empirical bayesian thresholding for sparse signals using mixture loss functions. Statistica Sinica 21: 449.

34. Johnson V (2008) Properties of bayes factors based on test statistics. Scandinavian Journal of Statistics 35: 354–368.

35. Varadhan R, Roland C (2008) Simple and globally convergent methods for accelerating the convergence of any em algorithm. Scandinavian Journal of Statistics 35: 335–353.

36. Storey J (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B (Statistical Methodology) 64: 479–498.

37. R Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`. Accessed June 3, 2013.

38. Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer New York.

39. Xie Y (2013) Dynamic Documents with R and knitr, volume 29. CRC Press.

Figure Legends

Tables

Supporting Information Legends

Supplementary material can be found in **Supplementary Information S1.**