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 improve inference. The first idea is to *assume that the distribution of effects is unimodal*, and to perform non-parametric inference under this assumption. This idea leads to a very simple, fast, and stable computer implementation, and improves inference of FDR, at least when the unimodal assumption is correct. (And when the assumption is incorrect, errors in inference are not too egregious.) 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 accounted for, and avoids the 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 terminology "signed discovery" for a discovery (an effect declared to be non-zero) for which the sign of the effect (positive or negative) is declared, and correspondingly define the "false signed discovery rate" (FSDR) in a natural way. We show that in some settings, particularly those with many discoveries, the FSDR and FDR can be quite different, and emphasise benefits of the FSDR, particularly its increased robustness to modelling assumptions.

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 espe-

cially 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).

Our methods are implemented in an R package, `ashr` (for **a**daptive **sh**rinkage in **R**), available at `http://github.com/stephens999/ashr`.

Methods

Notation

Suppose that we are interested in the values of J "effects" β_j ($j = 1, \ldots, J$). For example, in a canonical genomics application β_j might be the difference in the mean (log) expression of gene j in two conditions. In some contexts interest may focus on which of the β_j are "significantly" different from zero, whereas in other contexts interest may focus on estimating their values; the methods described here tackle both these problems. We assume that we have obtained data D that consist of estimates $\hat{\beta}_1, \ldots, \hat{\beta}_J$ of these effects, with corresponding (estimated) standard errors $\hat{s}_1, \ldots, \hat{s}_J$. Let H_j denote the event that the jth null hypothesis holds ($H_j : \beta_j = 0$), and t_j denote the t statistic testing this null:

$$t_j := \hat{\beta}_j / \hat{s}_j. \tag{1}$$

The t statistics, of course, have a t distribution under the null:

$$t_j | H_j \sim T_\nu \tag{2}$$

where T_ν denotes the t distribution on ν df, with ν assumed known. Let p_j denote the p value obtained by comparing t_j with its null distribution,

$$p_j := \Pr(|T_\nu| > |t_j|). \tag{3}$$

Some methods for FDR analysis require z scores instead of t statistics or p values; these methods can be applied by translating each t_j into a corresponding z score, which we denote z_j.

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:

$$\text{FDR}(\Gamma) := \frac{\#\{j : H_j \cap j \in \Gamma\}}{\#\{j : j \in \Gamma\}}, \tag{4}$$

taken to be 0 if both numerator and denominator are 0. Frequentist approaches to controlling the FDR attempt to control the expectation of FDR(Γ) 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 FDR(Γ). This posterior distribution can be used to obtain a point estimate for FDR(Γ), 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:

$$\text{lfdr}_j := \Pr(H_j|D). \tag{5}$$

The FDR can be estimated by the posterior mean for FDR(Γ), given by

$$\widehat{\text{FDR}}(\Gamma) := \frac{\sum_{j\in\Gamma} \text{lfdr}_j}{\#\{j : j \in \Gamma\}}. \tag{6}$$

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 [16], particularly his Theorem 1.

The q value for observation j is defined [16] 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

$$q_j := \widehat{\text{FDR}}(\{k : \text{lfdr}_k \leq \text{lfdr}_j\}). \tag{7}$$

The local False Sign Rate

[14] emphasise that there are settings where $\beta_j = 0$ is implausible, in which case the FDR is not useful: if every β_j is non-zero then there is no such thing as a false discovery and the FDR will always be identically 0. Gelman et. al. suggest that in such settings we should replace the concept of a false discovery with the concept of an "error in sign". The idea is that, in settings where $\beta_j = 0$ is implausible, the most fundamental inference objective is to ask which β_j are positive and which are negative. Indeed, even in settings where some β_j are exactly zero, separately identifying which are positive and which negative may be useful. For example, when identifying differentially expressed genes, analysts will often separate the genes that are "upregulated" from those that "downregulated" in a condition. Further, this idea has a long history: for example, [17], p32, suggests one should address

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

Motivated by this, we define the local False Sign Rate (lfsr) for observation j as

$$\text{lfsr}_j := \min[p(\beta_j \geq 0|\hat{\beta}, s), p(\beta_j \leq 0|\hat{\pi}, \hat{\beta}, s)]. \tag{8}$$

Thus lfsr_j gives the probability that we would get the sign of β_j wrong if we were to make our best guess. (Naturally, we consider it an error to state that β is positive or negative when it is truly zero.) To illustrate, suppose for concreteness that the minimum is achieved by the first term, $p(\beta_j \geq 0|\hat{\beta}, s) = 0.05$ say. Then our best guess would be that β is negative, and the probability that we have made an error in sign would be 0.05.

Note that $\text{lfsr}_j \geq \text{lfdr}_j$ because both the events $\beta_j \geq 0$ and $\beta_j \leq 0$ include the event $\beta_j = 0$. Thus, lfsr gives an upper bound for lfdr.

Because our methods involve explicitly modelling effects β_j, rather than p values or z scores, they can provide estimates of the local false sign rates. As show later, these estimates are much more robust to underlying assumptions than estimates of the FDR or lfdr. Thus, although for comparison with existing methods we have focussed on FDR computations, in practice we strongly recommend using lfsr and not lfdr.

Existing methods for FDR analysis

qvalue

One standard genomics analysis pipeline is to take the p values $p_1, \ldots, p_J$ and use them to estimate q values $q_1, \ldots, q_J$ using the R package `qvalue`. Figure 1 shows a graphical illustration of the procedure implemented in `qvalue`. In brief, the idea is that the overall distribution of p values can be decomposed into two constitutive components: one component corresponding to the "null" tests, and a second component corresponding to the "alternative" tests. 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 1). Specifically, at threshold γ, and given an estimate $\hat{\pi}_0$ for π_0, a natural non-parametric estimate of the FDR is

$$\hat{\text{FDR}}(\gamma) = J\hat{\pi}_0\gamma/\#\{j : p_j \leq \gamma\}. \tag{9}$$

See [18], equation (8).

EB approaches: `locfdr` and `mixfdr`

The fundamental idea of EB procedures is similar to that of `qvalue`: partition the observations into null and alternative components. Here we highlight two EB methods that seem most

relevant to our work here and which had good software implementations at the time of writing. The first is due to [6], implemented in the R package `locfdr`. This method differs from `qvalue` in that, instead of working with p values, it works with z scores. In the simplest case `locfdr` makes assumptions analogous to the two key assumptions of `qvalue`: i) the null z scores have a $N(0,1)$ distribution, and ii) all the z scores near 0 are from the null, which Efron calls the "Zero Assumption" (ZA). The `locfdr` package first estimates the overall distribution of z scores using a non-parametric regression approach, and then, using the two assumptions, "subtracts out" the null component, leaving the remainder to be the alternative component. In this way `locfdr`, like `qvalue`, avoids explicit modelling of the alternative component. (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.)

Another EB method, and perhaps the one most closely related to our work here, is due to [11], implemented in the R package `mixfdr`. Like `locfdr`, `mixfdr` works with z scores, but it differs in that it *explicitly* models the alternative distribution of z scores, using a mixture of normal distributions. In principle this approach also differs from `qvalue` and `locfdr` in that it does not explicitly make the ZA. However, we have found that in practice the results often approximately satisfy the ZA, which we believe to be due to details in the fitting procedure used.

Adaptive Shrinkage

Here we introduce an EB method, which, for reasons discussed above, we call "adaptive shrinkage" or "ash". The method provides not only estimated lfdr_j and q values, but also shrinkage point estimates and credible intervals for β_j, and estimates of the local false sign rates.

We begin by presenting the simplest version of our model, before embellishing it. Instead of partitioning the (observed) z scores or p values into null and alternative components, we use a model that partitions the (unobserved) effects β_j into null and alternative components. That is, letting g denote the distribution of the effects β_j, we write

$$g(\beta_j) = \pi_0 \delta_0 + (1 - \pi_0) g_1(\beta_j) \tag{10}$$

where δ_0 denotes a point mass at 0, and g_1 denotes the (unknown) distribution of the non-zero β_j. Thus π_0 denotes the proportion of effects that are null.

At this point we introduce a key assumption, which distinguishes our approach from previous work: we assume that g_1 is unimodal, with a mode at 0. One simple way to achieve this "Unimodal Assumption" (UA) is to assume that g_1 is a mixture of zero-mean normal distributions

$$g_1(\cdot) = \sum_{k=1}^{K} w_k N(\cdot; 0, \sigma_k^2) \tag{11}$$

where $N(\cdot; \mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and standard deviation σ. Here the w_k are non-negative mixture weights that sum to 1 and are to be estimated, and

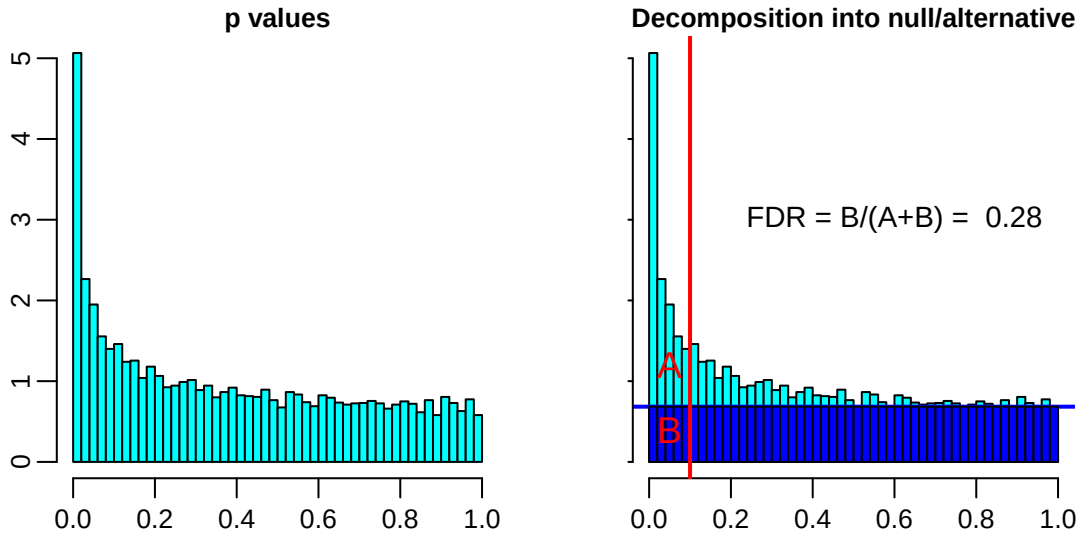


Figure 1. Illustration of how FDR is estimated by the `qvalue` package, by decomposing the overall distribution of p values (left) into a null component and alternative component (right). This is done by estimating the density of p values near 1, and drawing a horizontal line at that height to represent the density of the p values from the null distribution (based on the assumption that all p values near 1 are null). At a given threshold (red vertical line at 0.1), the area A corresponds to the true discoveries, and the area B corresponds to the true discoveries, so the FDR is estimated by $B/(A+B)$ which in this case is 0.28.

the 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). The symmetry of the normal distribution implies that (11) is not only unimodal, but also symmetric about zero. We discuss more flexible alternatives based on using mixtures of uniform distributions, and details (such as choice of gridpoints $\{\sigma_k : k = 1, \ldots, K\}$), as well as the UA more generally, below.

Since the β_j are unobserved, we need a model to connect the β_j to observations. Here we treat the estimates $\hat{\beta}_j$ as the observed data, and assume that they are noisy measurements of the true β_j with given standard error $\hat{s}_j$:

$$\hat{\beta}_j | \beta_j, \hat{s}_j \sim N(\beta_j, \hat{s}_j^2). \tag{12}$$

This approach, which follows [19], should be a good approximation if $\hat{\beta}_j$ and $\hat{s}_j$ are based on sufficiently many observations. We discuss below an improvement for moderate numbers of observations, based on replacing this normal likelihood with a t likelihood.

Combining (10)-(12), and assuming independence of the observations and effects for different j (given $\hat{s}, \pi_0, w_1, \ldots, w_k$), yields a hierarchical model for $\beta, \hat{\beta}$ given $\hat{s}$, with unknown hyper-parameters $(\pi_0, w_1, \ldots, w_k)$. To simplify notation we define $\pi_k := (1-\pi_0)w_k$ so that the unknown hyper-parameters can be written as a vector $\pi := (\pi_0, \pi_1, \ldots, \pi_K)$ whose elements are non-negative and sum to 1. The usual empirical Bayes approach to fitting this model would involve two steps:

1. Estimate the hyperparameters π by maximum likelihood, yielding a maximum likelihood estimate $\hat{\pi}$.

2. Compute quantities of interest (e.g. local fdr values, q values, credible intervals etc.) from the conditional distributions $p(\beta_j | \hat{\beta}_j, \hat{s}_j, \hat{\pi})$.

We make one modification to this procedure: instead of obtaining $\hat{\pi}$ by maximizing the likelihood, we instead maximise a penalized likelihood, where the penalty (detailed below) encourages $\hat{\pi}_0$ to be as big as possible (within the constrains of the model and observed data). We use this penalty here because in FDR appllications it is considered desirable to avoid underestimating π_0 so as to avoid underestimating the FDR. Both steps 1 and 2 are very straightforward: $\hat{\pi}$ can be obtained using a simple EM algorithm [20] to maximize the penalized likelihood, 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 (11), 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.)

The conditional distributions $p(\beta_j | \hat{\pi}, \hat{\beta}_j, \hat{s}_j)$ encapsulate uncertainty in β_j, combining information across *all* observations $j = 1, \ldots, J$. Specifically, the combining of the information occurs through estimation of π, which involves all of the data. Importantly, when estimating π our approach accounts for the precision of each measurement through the likelihood (12). For example, observations with a large $\hat{s}_j$ have an essentially flat likelihood, and so have little effect on estimation of π, and corresondingly little effect on the estimated FDR of other observations. We

highlight this because it contrasts with approaches that model the p values or z scores directly: observations with large $\hat{s}_j$ will tend to produce p values with approximately $U[0,1]$ distribution, and z scores with an approximately $N(0,1)$ distribution, under both the null and the alternative hypothesis (assuming that the effect β_j is small compared with $\hat{s}_j$). Such noisy observations therefore do affect inference; specifically they add to the estimated weight of the null component π_0, and inflate estimates of the FDR. Because of this our approach can take better account of the informativeness of each measurement, as illustrated in Results below.

Discussion on UA

Since we view the UA as a key assumption we briefly discuss it here. Although the UA will not apply to all situations, we argue that it will be reasonable in many practical contexts, especially those involving FDR where interest focusses on which β_j differ from zero. This is because whenever "$\beta_j = 0$" is a plausible null hypothesis, it seems reasonable to imagine that "β_j very near 0" is also plausible. From there it seems reasonable to imagine that larger effects become decreasingly plausible, and so the distribution of the effects will be unimodal about 0. Alternatively, we can 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. Note 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.

In addition, we note that although the UA is not made by any commonly-used existing FDR methods, 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 [21]. (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 [22] for work in this vein.) The UA assumption on regression coeffients 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.

Embellishments

More flexible unimodal distributions

To provide a more general approach to defining the distribution of effects $g(\beta_j)$, consider

$$g(\cdot;\pi) = \pi_0\delta_0(\cdot) + \sum_{k=1}^{K} \pi_k f_k(\cdot) \tag{13}$$

where f_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) $f_k(\cdot) = U[\cdot; -a_k, 0]$ 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 (13), so for a grid of values $a_1, \ldots, a_K$ there are actually $2K + 1$ components in the mixture.) The version of our model described above corresponds to (i). Replacing these with uniform components (ii)-(iii) only slightly complicates calculations under the normal likelihood (12), and greatly simplifies the calculations under the t likelihood (14) introduced below. Our use of uniform components here closely mirrors [23].

Moving from (i) to (iii) the representation (13) 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 [24] and [25]; see [26], 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 **??** for our software defaults.

Replace normal likelihood with t likelihood

We can generalize the normal likelihood assumption (12) by replacing it with a t likelihood:

$$\hat{\beta}_j \,|\, \beta_j, \hat{s}_j \sim T_\nu(\beta_j, \hat{s}_j) \tag{14}$$

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 (12) corresponds to the limit $\nu \to \infty$. This generalization does not complicate inference when used with mixtures of uniform components for g in (13). For normal components the computations with a t likelihood are considerably more difficult and we have not implemented this combination.

Equation (14) is, of course, motivated by the standard asymptotic result

$$(\hat{\beta}_j - \beta_j)/\hat{s}_j \sim T_\nu. \tag{15}$$

However (15) does not imply (14), because in (15) $\hat{s}_j$ is random whereas in (14) 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).

Dependence of effects on standard errors

Equation (10) 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

$$\frac{\beta_j}{\hat{s}_j^\alpha} \,|\hat{s}_j \sim g(\cdot) \tag{16}$$

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 [19] 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 16 for general α is actually easily 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 $\beta_j = b_j\hat{s}_j^\alpha$. In the special case $\alpha = 1$, $\hat{b}_j$ is the t score t_j and $\hat{s}'b = 1$, so this involves applying our algorithm to the t scores treating them all as having the same standard error. In this case the local fdr from our method becomes monotonic in each direction as the t score increases away from zero, just as with `qvalue`, and (usually) `mixfdr` and `locfdr`.

Different values of α in (16) 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 (16) for any α, this approach seems preferable in practice than simply assuming $\alpha = 0$ or $\alpha = 1$. For brevity we do not investigate this idea any further here, but it has been implemented in our software by C.Dai, using a simple grid search over α.

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 **??** for further implementation details.

Results

Effects of the Unimodal Assumption

Here we consider the effects of making the UA. To isolate these effects we consider the simplest case, where every observation has the same standard error, $s_j = 1$, and we consider the normal limit (**??**). That is,

$$\hat{\beta}_j|\beta_j \sim N(\beta_j, s_j^2 = 1) \tag{17}$$

so $z_j := \hat{\beta}_j/s_j = \hat{\beta}_j$ is standard normal if null H_j holds.

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 nulls that hold, π_0, and hence conservative estimates of FDR and lfdr.
3. The UA results in a stable procedure, both numerically and statistically.
4. The UA is somewhat robust to deviations from unimodality.

0.0.1 The UA compared with the ZA

Figure 2 graphically illustrates the difference between the ZA made by existing methods, and the UA made by our method. The effects of the ZA are visually clear, producing a "hole" in the alternative z score distribution around 0 for `qvalue`, `locfdr` and `mixfdr`; in contrast, the UA fills in this hole to produce a peak at 0. 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") as one would hope.

How do these differences influence estimates of the FDR? To address this, we note that the primary driver of differences in FDR estimates among methods is *differences in their estimates of* π_0, the proportion of nulls (at least in this simple case where the s_i are all identical). To see why, observe that, given an estimate $\hat{\pi}_0$ for π_0, equation (9) provides a strikingly natural non-parametric estimate for the FDR (provided the denominator is not very small). Thus, although the methods considered here differ from one another in many ways, having estimated π_0 they should, if sensible, produce estimates of FDR that agree closely with (9). And, indeed, our limited empirical checks suggest that this is indeed the case (`https://www.github.com/stephens999/ash/examples/qvalue.nonp.rmd`).

Thus, differences in FDR estimates among methods are driven primarily by differences in their estimates of π_0, and so to assess differences in FDR estimates among methods we focus on assessing differences in estimates of π_0. Figure 2 suggests that, compared with the ZA, the UA will tend to reduce the estimate of π_0. At the same time, provided the UA holds, we hope still to provide "conservative" estimates of π_0 (that is, overestimates of π_0, which will yield overestimates of the FDR).

To confirm this we performed comprehensive simulation experiments across a range of alternative scenarios (i.e. values for the distribution of non-zero effects, which we denote g_1), and values for π_0. The alternative distributions are shown in Figure **??**a, with details in Table 1.

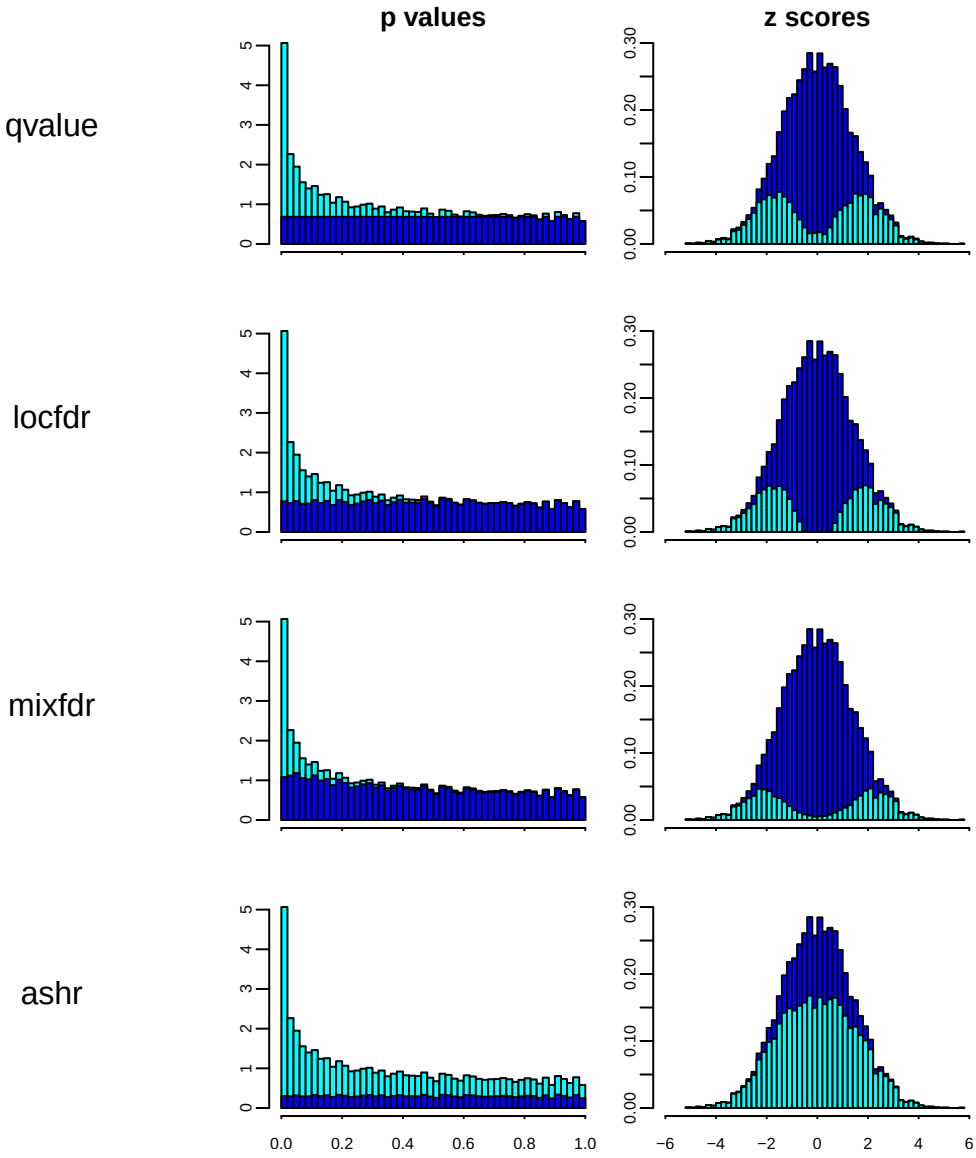


Figure 2. Comparison of the way that different methods decompose p values (left) and z scores (right) into a null component (dark blue) and an alternative component (light blue). In the z score space the alternative distribution is placed on the bottom to highlight the differences in its shape among 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, $\beta_j \sim N(0, 1^2)$, $s_j = 1$, resulting in $z_j \sim N(0, 2^2)$ – so the "true" decomposition would make everything light blue.)

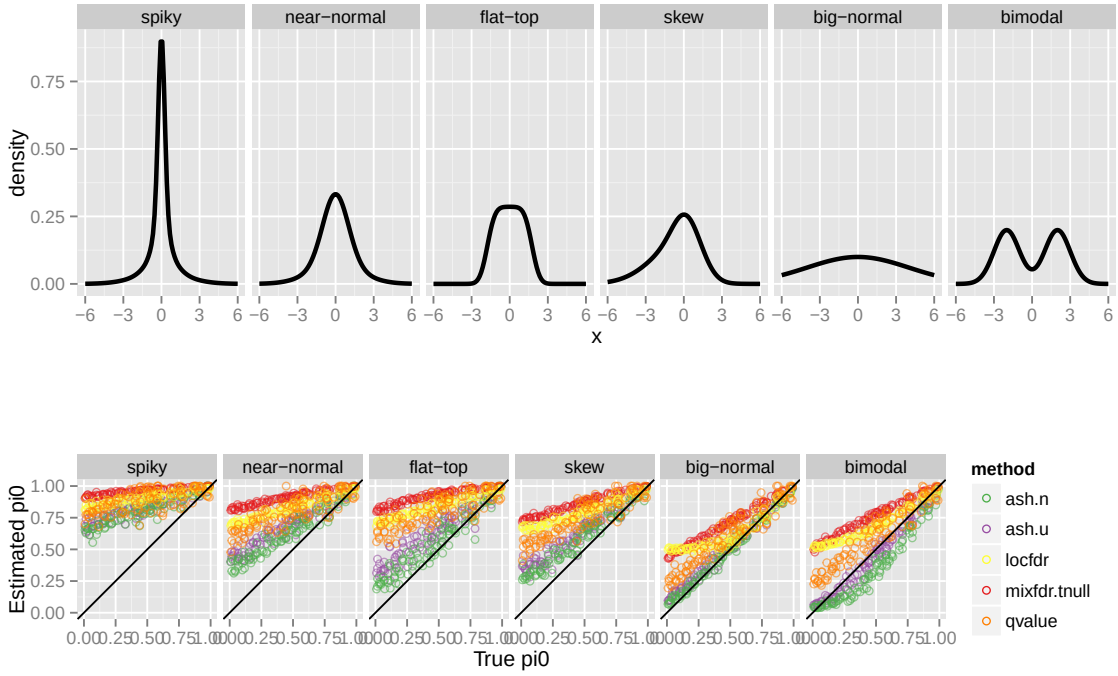


Figure 3. a) Summary of the alternative distributions used for different simulation scenarios. b) Comparison of true and estimated values of π_0 in these simulated scenarios. In general π_0 is impossible to estimate accurately, and all methods yield conservative (over-)estimates for π_0 (deliberately so, to ensure conservative estimates of FDR). As expected, when the UA holds, the `ashr` method is least conservative, and hence most accurate. When the UA assumption does not hold ("bimodal" scenario) the `ashr` estimates are slightly anti-conservative, but quite accurate, particularly when using a mixture of uniforms to model the unimodal distribution.

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 much 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 our methods 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 $\beta_j \sim \pi_0\delta_0 + (1 - \pi_0)g_1(\cdot)$.

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 1. Summary of simulation scenarios considered

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

Thus these simulations assume the same precision ($s_j = 1$) for each measurement.

We applied the methods implemented in the R packages `qvalue`, `locfdr`, `mixfdr` and `ashr` to estimate π_0 for each data set. For `ashr` we show results for g modelled as a mixture of normal components ("ash.n") and 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.)

Figure **??** compares estimated and true values of π_0 under each Scenario. For most scenarios no method reliably estimates π_0. This is expected because most scenarios include an appreciable fraction of "small non-null effects" that are essentially indistinguishable from 0, making accurate estimation ofπ_0 impossible. This is particularly true for the "spiky" scenario, for example. The exception is the "big-normal" scenario, where the `ashr` estimates of π_0 are quite accurate - made possible by the fact that, in this scenario, most non-zero effects are very different from zero.

Although accurate estimation of π_0 is impossible in general, the simulation results demonstrate that, when the UA holds, `ashr` consistently provides substantially smaller, more accurate, estimates of π_0, than existing methods. As noted above, this means that estimates of FDR from `ashr` will also be smaller and more accurate under these scenarios (while remaining conservative). When the UA assumption does not hold ("bimodal" scenario) the `ashr` estimates are slightly anti-conservative, although quite accurate. We view such a strong bimodal scenario as rather unlikely in most applications where FDR methods are used; nonetheless, users may find it reassuring that even in this case the results from `ashr` should not be horribly misleading.

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 (XX need to specify in methods?). We then compared the largest log-likelihood achieved across all 11 runs with the log-likelihood achieved by our 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 address any question about the effect sizes, including, for example, "which effects likely exceed T", where T is some effect threshold of interest (and "likely" can be quantified by whatever probability threshold is desired). Importantly, because the posterior distribution for β_j depends on the "prior" effect-size distribution g, which is estimated from *all* the observations, this posterior reflects information from all the observations and not only from the observation $\hat{\beta}_j$. In our case, because we assume g to be unimodal about zero, this prior always "shrinks" the posterior distributions towards zero, with the strength of the shrinkage being estimated from the data. Further, because the posterior distribution is, by definition, conditional on the observed data, inferences 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 shrunken posterior distributions and interval estimates for parameters that are valid "post-selection" 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 in constructing the empirical Bayes Fdr estimates, but it is hard to imagine a generally satisfactory parametric formulation..." ([6] rejoinder, p46). The key point we want to emphasise here is that *if we restrict ourselves to settings where we can assume that g is unimodal* then a stable non-parametric approach becomes not only possible, but indeed quite straightforward. While not meeting Efron's desire for an entirely general nonparametric approach, we believe that these settings include many cases of practical interest.

To illustrate this, Figure **??** compares the estimated g from our method for each of the simulation scenarios above with the estimated g from `mixfdr` which does not make the UA (and which models g as a mixture of J normal distributions, with $J = 3$ by default). Even with only

one simulated example for each scenario, the greater reliability of estimates afforded by the UA is immediately apparent. In particular the greater tendency for `mixfdr` to overestimate π_0 is readily visible in its behaviour at 0. And the estimated cdf from `mixfdr` often has a vertical (or almost vertical) segment at some non-zero location, indicative of a concentration of density in the estimated g at that location. This corresponds to one of the 3 mixture components being estimated to be centered at that location with very small variance. 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.

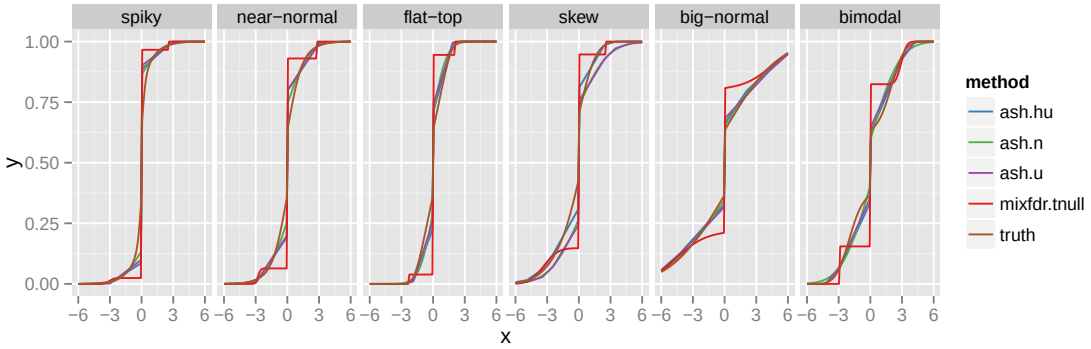
Interestingly, even in the "bimodal" scenario where the UA is severely contradicted, the estimated g from `ashr` is visually more accurate than that from `mixfdr`: although `mixfdr` is capable, in principle, of fitting the multiple modes of g, it does not do this well here. We believe this mostly reflects the fact that the noise level is sufficiently large to make reliable estimation of the multiple modes difficult, rather than fundamental deficiencies of `mixfdr`. Indeed, in multimodal 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 multimodality is clearly visible in the observed $\hat{\beta}$.

We note one caveat on the accuracy of estimated g: we tend to systematically overestimate the mass of g near zero, due to our deliberate use of a penalty term (**??**) to avoid underestimation of π_0. On careful inspection, this is visually apparent in Figure **??**: 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 **??**b), and applying our methods without the penalty term removes this systematic effect, although at the cost of sometimes under-estimating π_0 (Figure **??**c).

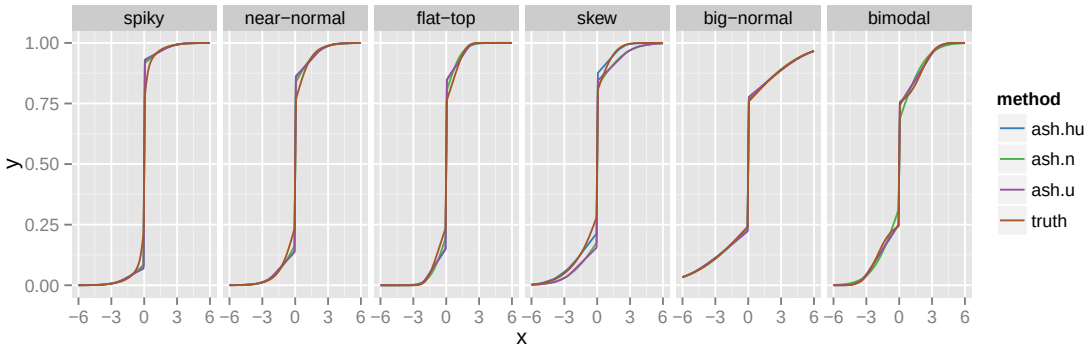
Calibration of posterior distributions

To quantify the effects of errors in estimates of g we examine the calibration of the resulting posterior distributions (averaged over all 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 asymetric scenario). We 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.

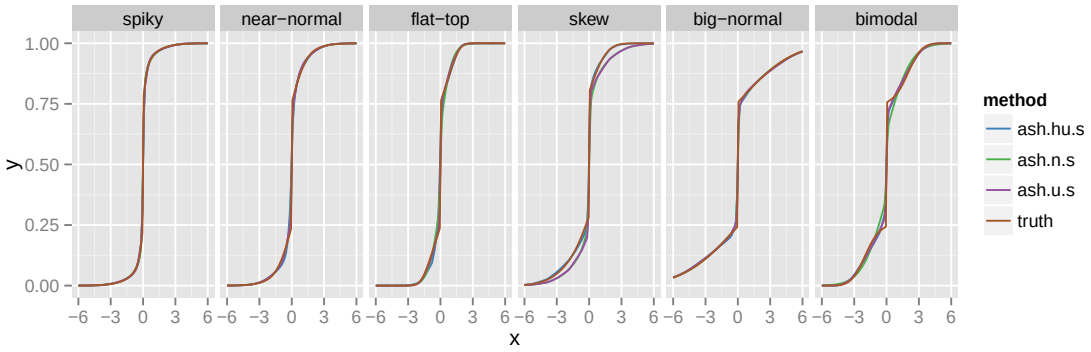
The results are shown in Table 2. As noted above, our use of a penalty term to avoid underestimating π_0 means that the estimated gs tends to have, systematically, slightly too much mass near 0. If the prior has too much mass near 0 then so will the posterior, so we should expect posterior intervals may also be "over-shrunk" towards zero. The effect of overshrinking on the left tail area is different for "negative discoveries" and "positive discoveries": for negative discoveries the lower credible bound will be negative and overshrinking will move the bound



(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 (**??**) so π_0 is systematically overestimated.



(c) Average estimated cdfs across ~ 10 data sets compared with truth; methods here do not use penalty (**??**) 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 4. Comparisons of estimated cdfs of g and true cdf of g. See Figure **??** for simulation scenarios.

towards zero, making the left tail too big; for positive discoveries, the lower confidence bound will be positive, and overshrinking will move it toward zero, making the tail too small. In the context of reliably identifying "large" effects we might worry more about under-shrinkage, and be prepared to put up with some overshrinkage to avoid it.

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

When we tried using no penalty term in the half-uniform case we came across an unanticipated problem: 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.

Note - this could probably be solved by making the small components symetric - the idea is that the asymetry only comes in with the tail.

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 our use of a likelihood, (12) or (14), which explicitly incorporates the measurement precision (standard error) of each $\hat{\beta}_j$.

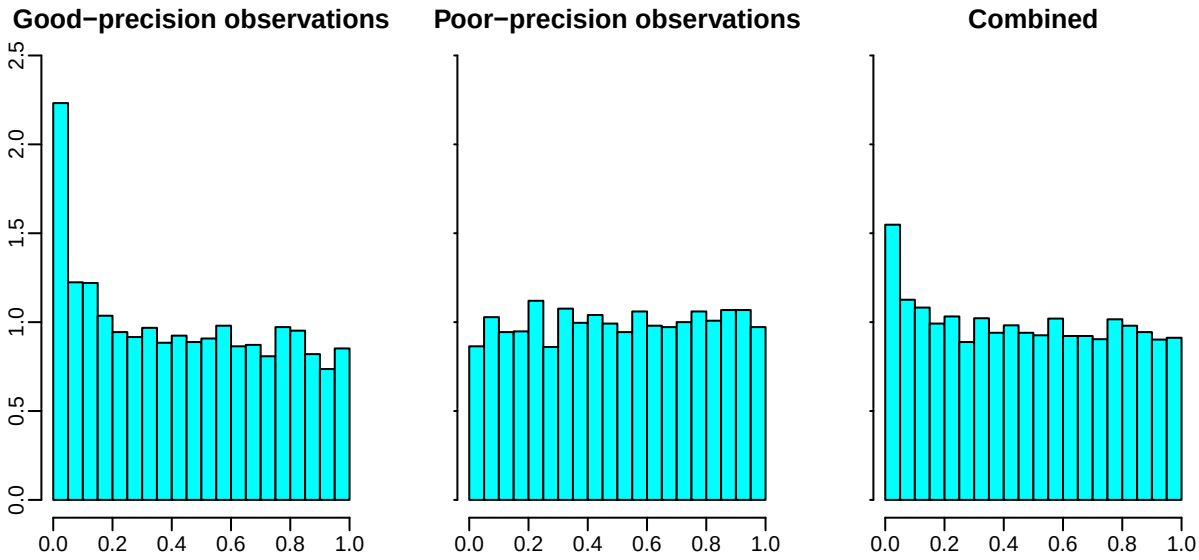
To illustrate, we begin with a simple simulation where half the measurements are quite precise (standard error $s_j = 1$), and the other half are very poor ($s_j = 10$). In both cases, we assume that half the effects are null and the other half are normally distributed with standard deviation 1:

$$p(\beta) = 0.5\delta_0(\beta) + 0.5N(\beta; 0, 1). \tag{18}$$

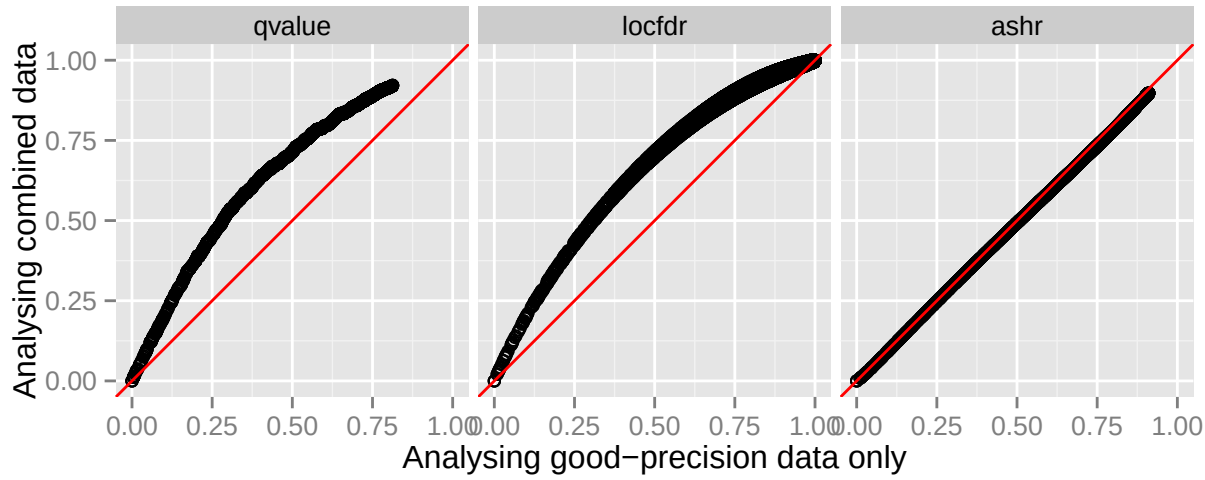
In this setting, the poor-precision measurements tell us very little, and any sane analysis should effectively ignore them. However, this is not the case in standard FDR-type analyses (Figure **??**). This is because the poor measurements produce p values that are approximately uniform (Figure 5a), which, 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 5b). In contrast, the results from `ashr` for the good-precision observations are unaffected by including the low-precision observations in the analysis.

Adaptive Shrinkage

[27] note the ability of Empirical Bayes 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 distribution of effects (g in (**??**)) 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



(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 estimated error rate (q values from `qvalue`; lfdr for `locfdr`; lfsr for `ashr`) for a good-precision observation under the two different analyses.

Figure 5. Simulation illustrating how 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.

	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 **??** 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. (Removing the penalty in the half-uniform case can lead to unexpected problems explained in section **??**).

	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 2. Table of empirical coverage for nominal 95% lower credible bounds (methods with penalty term).

also tend to concentrate around zero (strong shrinkage). In contrast, if most β_j are large then $\hat{g}$ will reflect this by being 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 non-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, (12) or (14), 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

	spiky	near-normal	flat-top	skew	big-normal	bimodal
ash.n.s	0.95	0.95	0.95	0.95	0.96	0.96
ash.u.s	0.94	0.95	0.95	0.94	0.96	0.96
ash.hu.s	0.88	0.92	0.92	0.92	0.92	0.93

(a) All observations

	spiky	near-normal	flat-top	skew	big-normal	bimodal
ash.n.s	0.95	0.95	0.99	0.93	0.95	0.97
ash.u.s	0.89	0.91	0.89	0.91	0.94	0.94
ash.hu.s	0.89	0.92	0.91	0.94	0.95	0.94

(b) "Significant" negative discoveries.

	spiky	near-normal	flat-top	skew	big-normal	bimodal
ash.n.s	0.94	0.94	0.92	0.88	0.95	0.94
ash.u.s	0.94	0.93	0.93	0.88	0.95	0.95
ash.hu.s	0.32	0.61	0.53	0.54	0.79	0.82

(c) "Significant" positive discoveries.

Table 3. Table of empirical coverage for nominal 95% lower credible bounds (methods without the penalty term.

6 illustrates the idea by contrasting results for two simulation scenarios with different signal strengths, where observations vary in their measurment precision.

Model comparison, and the p value prior

Thus far we have focussed on the model (**??**), where the distribution of β_j is independent of the standard error s_j. This model corresponds to the case $\alpha = 0$ in the more general model (16). In the previous section we emphasised that, under this model, imprecise measurements are shrunk more strongly than precise measurements, and that the lfsr and lfdr are not a monotonic function of p value. In this section we emphasise that this behaviour *depends on the assumption that β_j is independent of the standard error s_j.* Specifically, if we make a different assumption, setting $\alpha = 1$ in (16) so β_j scales linearly with s_j, then we get different behaviour, as summarized by the following result.

Result 1 *Treat the standard errors s_j as fixed and given. Under the normal likelihood (12), and under any prior distribution of the form (16) with $\alpha = 1$, the (Bayesian) lfsr$_j$ and lfdr$_j$ are all monotonic with $|t_j|$ where $t_j = \hat{\beta}_j/s_j$, and with the corresponding p value (p_j). Also, the shrinkage factor (Posterior Mean(β_j)/$\hat{\beta}_j$) is montonic with $|t_j|$.*

This result is a translation and extension of the result of [19], who showed that, when testing a series of null hypotheses, with a normal likelihood, Bayes Factors and p values produce the

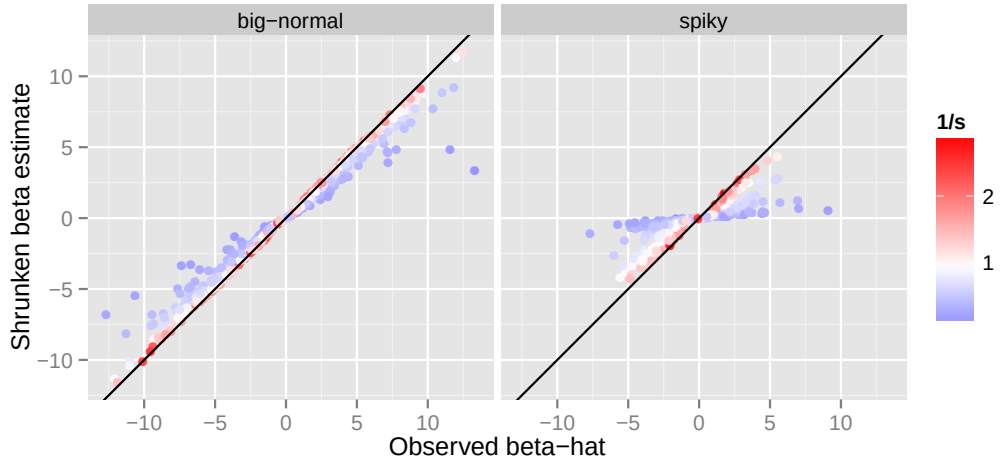


Figure 6. Figure illustrating adaptive shrinkage. Results are shown for two different scenarios: (left) "big-normal" where the effects have a wide distribution; (right) "spiky", where the effects are more concentrated near 0. The standard error of each observation, s_j, was simulated from Inverse Gamma(5,5), 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.

same ranking of tests *if* the Bayes Factor for the jth test is computed using a normal prior with mean 0 and variance Ks_j^2 for some constant K.

Most previous authors of empirical Bayes approaches to false discovery rates [6, 11] have worked with models for z scores β_j/s_j, rather than models for the effects β_j. In doing this they assume (implicitly or explicitly) that the z scores are identically distributed under the alternative model. This is analogous to assuming $\alpha = 1$ in (refeqn:beta-alpha). And, given the result above, standard frequentist FDR methods, like `qvalue` and Benjamini–Hochberg, that rank tests by their p value, could also be viewed as making the same implicit assumption. In this sense, setting $\alpha = 1$ in our approach provides the closest correspondance with existing methods (although the unimodal assumption still distinguishes our approach from existing methods)

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 - 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 choose whichever of $\alpha = 0, 1$ yields the highest likelihood, and this is also straightforward.

previous work based on normal means: http://link.springer.com/article/10.1007/BF02562623

Local False Sign Rate

There are two reasons to use the lfsr instead of the lfdr: it is more generally meaningful (e.g. it applies whether or not zero effects exist), and estimation of lfsr is more robust to modeling assumptions and estimation of π_0. To illustrate this, we compared estimated and true values of both lfdr and lfsr for the simulated data (where the true values are computed by Bayes Theorem using the true value of g).

0.1 Null behaviour

Results of simulations under the null.

Newcombe (1886) Dean and Raftery

0.2 Do we need a point mass at zero?

In some settings it is the convention to focus on testing whether $\beta_j = 0$. However some dislike this focus, objecting that it is unlikely to be the case that $\beta_j = 0$ exactly. For example, when comparing the average expression of a gene in human samples vs chimp samples, it might be considered unlikely that the expression is *exactly* the same in both. Whether or not $\beta_j = 0$ is considered unlikely may depend on the context. However, in most contexts, finite data cannot distinguish between $\beta_j = 0$ and β_j being very close to zero. Thus finite data cannot usually convince a skeptic that β_j is exactly zero, rather than just very small. In contrast it is easy to imagine data that would convince a doubter that β_j is truly non-zero. In this sense there is an assymetry between the inferences "β_j is zero" and "β_j is non-zero", an assymetry that is reflected in the admonition "failure to reject the null does not imply it to be true".

Thus any analysis that purports to distinguish between these cases must be making an assumption.

Consider two analyses of the same data, using two different "priors" g for β_j, that effectively differ only in their assumptions about whether or not β_j can be exactly zero. For concreteness, consider

$$g_1(\cdot) = \pi\delta_0(\cdot) + (1-\pi)N(\cdot; 0, \sigma^2)$$

and

$$g_2(\cdot) = \pi N(\cdot; 0, \epsilon^2) + (1-\pi)N(\cdot; 0, \sigma^2).$$

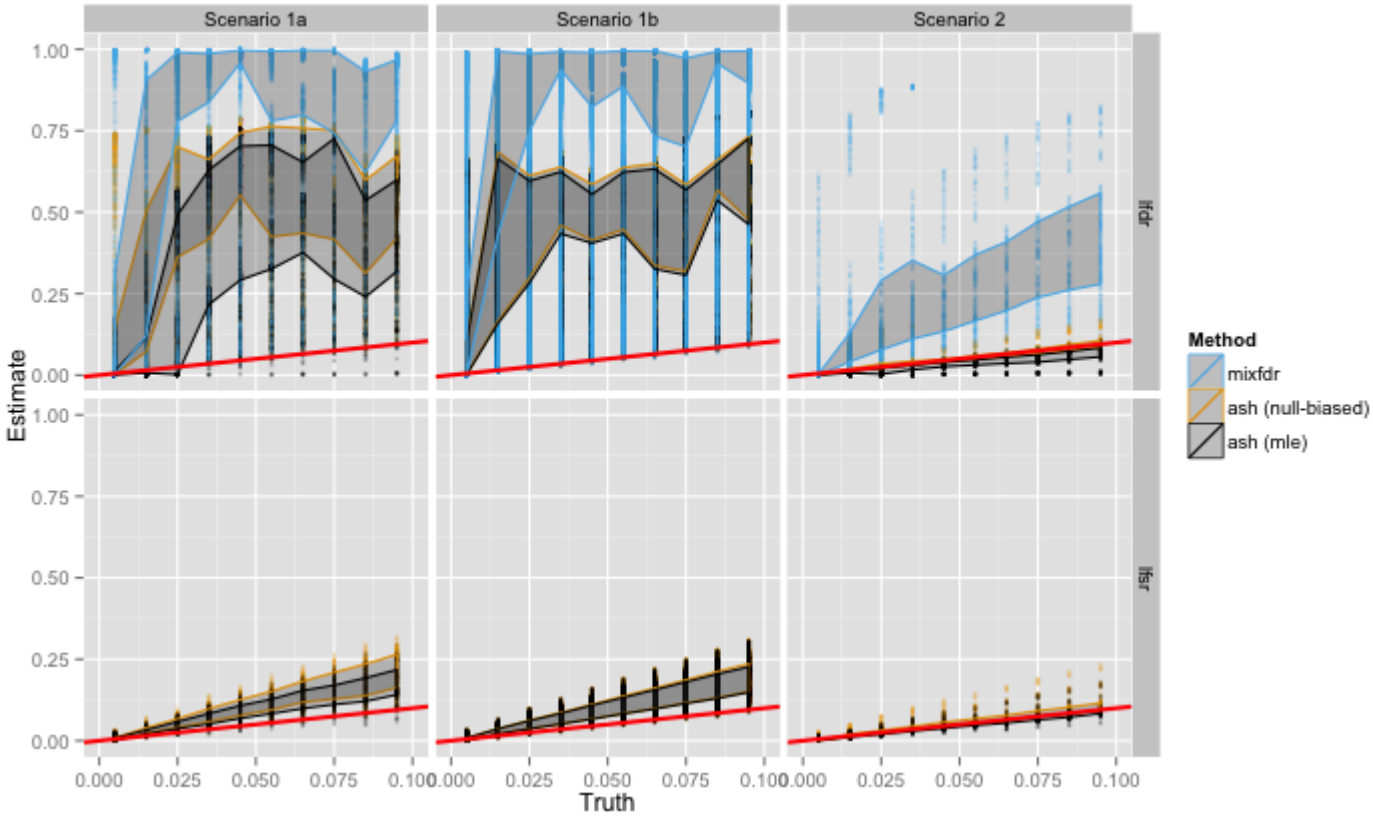


Figure 7. Figure showing lfsr is more robust than lfdr

If ϵ^2 is sufficiently small, then these priors are "approximately the same", and will lead to "approximately the same" posteriors and inferences in many senses. To discuss these, let p_j denote the posterior under prior g_j. Then, for any given (small) δ, we will have $p_1(|\beta_j| < \delta) \approx p_2(|\beta_j| < \delta)$. However, we will not have $p_1(\beta_j = 0) \approx p_2(\beta_j = 0)$: the latter will always be zero, while the former could be appreciable.

What if, instead, we examine $p_1(\beta_j > 0)$ and $p_2(\beta_j > 0)$? Again, these will differ. If this probability is big in the first analysis, say $1 - \alpha$ with α small, then it could be as big as $1 - \alpha/2$ in the second analysis. This is because if $p_1(\beta_j > 0) = 1 - \alpha$, then $p_1(\beta_j = 0)$ will often be close to α, so for small ϵ $p_2(\beta_j)$ will have mass α near 0, of which half will be positive and half will be negative. Thus if we do an analysis without a point mass, but allow for mass near 0, then we may predict what the results would have been if we had used a point mass.

Let's try: "'r beta.ash.pm = ash(ss$betahat$, ssbetasd, usePointMass=TRUE) print(beta.ash.pm) print(beta.ash.auto) plot(beta.ash.auto$localfsr$, $beta.ash.pm$localfsr,main="comparison of ash localfsr, with and without point mass",xlab="no point mass", ylab="with point mass",xlim=c(0,1),ylim=c(0,1) abline(a=0,b=1) abline(a=0,b=2) "'

Our conclusion: if we simulate data with a point mass, and we analyze it without a point

mass, we may underestimate the lfsr by a factor of 2. Therefore, to be conservative, we might prefer to analyze the data allowing for the point mass, or, if analyzed without a point mass, multiply estimated false sign rates by 2. In fact the latter might be preferable: even if we analyze the data with a point mass, there is going to be some unidentifiability that means estimating the pi value on the point mass will be somewhat unreliable, and we still might underestimate the false sign rate if we rely on that estimate. TO THINK ABOUT: does multiplying the smaller of Pr(¡0) and Pr(¿0) by 2, and adding to Pr(=0) solve the problem in either case?

0.3 Discussion

Cite johnston and silverman; raykar and zhao, sinica.

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

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 [19] (see also [28]) 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 [16], 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:

$$h(\pi; \lambda) = \prod_{k=0}^{K} \pi_k^{\lambda_k - 1} \tag{19}$$

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.

Initialization

We initialize our EM algorithm with $\pi_k = 1/n$ for $k = 1, \ldots, K$, with $\pi_0 = 1 - \pi_1 - \cdots - \pi_K$. 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.

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

$$p(\beta_j \mid \hat{s}, \pi) = \sum_{k=0}^{K} \pi_k f_k(\beta_j) \tag{20}$$

$$p(\hat{\beta}_j \mid \beta_j, \hat{s}, \pi) = N(\hat{\beta}_j; \beta_j, \hat{s}_j^2), \tag{21}$$

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

Multiplying (20) and (21) together gives $p(\hat{\beta}_j, \beta_j|\hat{s}, \pi)$, which integrating over β_j yields

$$p(\hat{\beta}_j|\hat{s}, \pi) = \sum_{k=0}^{K} \pi_k \tilde{f}_k(\hat{\beta}_j) \tag{22}$$

where

$$\tilde{f}_k(\hat{\beta}_j) := \int f_k(\beta_j) N(\hat{\beta}_j; \beta_j, \hat{s}_j^2)\, d\beta_j \tag{23}$$

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,

$$\tilde{f}_k(\hat{\beta}_j) = \begin{cases} N(\hat{\beta}_j; 0, \hat{s}_j^2 + \sigma_k^2) & \text{if } f_k(\cdot) = N(\cdot; 0, \sigma_k^2), \\ \frac{\Psi((\hat{\beta}_j - a_k)/\hat{s}_j) - \Psi((\hat{\beta}_j - b_k)/\hat{s}_j)}{b_k - a_k} & \text{if } f_k(\cdot) = U(\cdot; a_k, b_k), \end{cases} \tag{24}$$

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:

$$L(\pi) = \prod_j \sum_k \pi_k w_{kj} \tag{25}$$

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:

$$w_{kj} = \pi_k l_{kj} / \sum_{k'} \pi_{k'} l_{k'j} \tag{26}$$

$$n_k = \sum_j w_{kj} + \lambda_k - 1 \quad \text{E Step} \qquad \pi_k = n_k / \sum_{k'} n_{k'} \quad \text{M step} \tag{27}$$

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` [29]) to accelerate convergence of this EM algorithm.

Conditional distributions

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

$$p(\beta_j|\hat{\pi}, \hat{\beta}, s) \propto g(\beta_j; \pi) L(\beta_j; \hat{\beta}_j, \hat{s}_j). \tag{28}$$

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

$$p(\beta_j|\hat{\pi}, \hat{\beta}, s) = \sum_{k=0}^{K} w_{kj} p_k(\beta_j|\hat{\beta}_j, \hat{s}_j) \tag{29}$$

where the posterior weights w_{kj} are computed as in (26) 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.

Acknowledgements

Statistical analyses were conducted in the R programming language [30], Figures produced using the ggplot2 package [31], and text prepared using LaTeX. Development of the methods in this paper was greatly enhanced by the use of the `knitr` package [32] 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. Storey J (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. The Annals of Statistics 31: 2013–2035.

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

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

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

20. 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.

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

22. 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.

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

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

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

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

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

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

29. 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.

30. 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.

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

32. 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.**