Adaptive shrinkage and the false sign rate

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

It is common in statistics that you measure many "similar" things imperfectly, and wish to estimate their values. The situation arises commonly in the kinds of genomics applications I am often involved in, but also in other areas of statistics. In genomics, for example, a very common goal is to compare the mean expression (activity) level of many genes in two conditions. Let μ_j^0 and μ_j^1 denote the mean expression of gene j ($j = 1, \dots, J$) in the two conditions, and define $\beta_j := \mu_j^0 - \mu_j^1$ to be the difference. Typically expression measurements are made on only a small number of samples in each condition - sometimes as few as one sample in each condition. Thus the error in estimates of μ_j^0 and μ_j^1 is appreciable, and the error in estimates of β_j still greater.

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.

Suppose that we measure, with error, a series of "effects", $\beta_1, \dots, \beta_J$. To take just one concrete example, β_j could be the difference in the mean (log) expression levels of gene j ($j = 1, \dots, J$) between 2 conditions. In this case, a measurement of the effect might be the difference in sample means obtained in the two conditions. We will let $\hat{\beta}_j$ denote the measured value of β_j, and assume that each measurement comes with an associated standard error, s_j. A key aim here will be to take proper account of the fact that some measurements may be more precise than others: that is, to take proper account of variation in s_j across j.

A common goal, particularly in genomic studies, is to identify which β_j differ from zero. This is commonly tackled by first computing an effect size estimate ($\hat{\beta}_j$) and its standard error (s_j), converting this to a Z score ($Z_j = \hat{\beta}_j / s_j$) and a corresponding p value (p_j), testing $H_j : \beta_j = 0$. Then standard methods (e.g. the qvalue package) can be used to estimate False Discovery Rates at any given threshold.

There are two issues with this approach that I would like to address here. The first is that it really does not take proper account of the measurement errors. To see this, consider an example where half the measurements are quite precise, and the other half are really, really, poor. Intuitively, the poor measurements tell us nothing, and any sane analysis should effectively ignore them. However, in a standard FDR-type analysis, these poor measurements add "noise" and affect estimated FDRs. This is because the p values from the poor measurements will be effectively uniformly distributed, and some will be significant at any given threshold.

The second issue is that directly modeling the p values, say via non-parametric methods, without taking account of their precision, 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.

The initial goal of the ASH (Adaptive SHrinkage) project is to provide simple, generic, and flexible methods to derive "shrinkage-based" estimates and credible intervals for unknown quantities $\beta = (\beta_1, \dots, \beta_J)$, given only estimates of those quantities ($\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_J)$) and their corresponding estimated standard errors ($s = (s_1, \dots, s_J)$).

Although shrinkage-based estimation can be motivated in various ways, our key goal here is to combine information across the multiple measurements $j = 1, \dots, J$ to improve inference for each individual β_j. By improved inference, we mean both improved average accuracy of point estimates, which is the traditional focus of shrinkage-based methods, *and* improved assessments of uncertainty.

By "adaptive" shrinkage we have two key properties in mind. First, the appropriate amount of shrinkage is determined from the data, rather than being pre-specified. Second, the amount of shrinkage undergone by each $\hat{\beta}_j$ will depend on the standard error s_j: measurements with high standard error will undergo more shrinkage than measurements with low standard error.

Given that shrinkage estimation is widely recognized as a powerful tool, there are surprisingly few software packages for performing the simplest type of shrinkage estimation considered here. (There are more packages for the more complex setting of covariance estimation, where shrinkage is perhaps still more important.) The only package we have found that provides anything similar to the functionality provided here is mixfdr (Muralidharan). Compared with mixfdr, the key features of ashr are that it i) focuses on allowing for variation in the standard deviation of each observation; ii) constrains the underlying density to be unimodal (and possibly symmetric). NOTE: should emphasize these differences in the examples.

As an important special case, these methods address the "multiple comparisons" setting, where interest usually focuses on which β_j can be confidently inferred to be non-zero. Such problems are usually tackled by computing a p value for each j, often by applying a t test to $\hat{\beta}_j/s_j$, and then applying a generic procedure, such as that of Benjamini and Hochberg (1995?) or Storey (2001?), designed to control or estimate the false discovery rate (FDR) or the positive FDR (Storey, 2001?). In essence we aim to provide analagous generic methods that work directly with two numbers for each measurement $(\hat{\beta}_j, s_j)$, rather than a single number (e.g. the p value, or t statistic). Working with these two numbers has two important benefits: first, it permits estimation and not only testing; second, the uncertainty in each measurement $\hat{\beta}_j$ can be more fully accounted for, reducing the impact of "high-noise" measurements (large s_j) that can reduce the effectiveness of a standard FDR analysis.

The potential for shrinkage-based estimation to address the multiple comparisons setting has been highlighted previously, including Greenland and Robins (1991), Efron (2008) and Gelman et al (2012). [Note, check also Louis, JASA, 1984]

Methods

Suppose that we are interested in the values of n "effects" β_j $(j = 1, \dots, n)$. In some contexts our interest may focus on which of the β_j are "significantly" different from zero, whereas in other contexts our interest may focus on estimating their values; the methods described here are suited to both these contexts. We assume that we have obtained data $D_1, \dots, D_n$ that provide independent estimates $\hat{\beta}_1, \dots, \hat{\beta}_n$ of these effects, with corresponding (estimated) standard errors $s_1, \dots, s_n$. For our purposes, these estimates and standard errors could be obtained from the data using any method, provided they (approximately) satisfy

$$\hat{\beta}_j | \beta_j, s_j \sim N(\beta_j, s_j). \tag{1}$$

[In some settings a t distribution assumption for $\hat{\beta}_j | s_j, \beta_j$ may be more appropriate, and could also be incorporated into our methods; see Discussion.]

In outline, we use a hierarchical model to combine information across measurements, with the aim of improving both accuracy and precision of estimates. Specifically we assume that the effects β_j are independent and identically distributed from some unknown distribution $g(\cdot)$,

$$\beta_j \sim g(\cdot), \tag{2}$$

where g will be estimated from the data. Our methods make the following key assumptions, which distinguish them from most previous work in this area:

A1: The distribution $g(\cdot)$ is unimodal. (We assume here that the mode is at 0, although this could be relaxed.)

A2: The likelihood $L(\beta_j) := p(D_j|\beta_j)$ can be approximated by a normal likelihood,

$$\hat{L}_j(\beta_j) \propto \exp[-0.5(\beta_j - \hat{\beta}_j)^2/s_j^2]. \tag{3}$$

Discussion on Assumptions

Although the unimodal assumption, A1, will not apply to all situations, we argue that it will be reasonable in many practical contexts. For example, in contexts where interest focusses on which β_j differ from zero, which suggests that "$\beta_j = 0$" is a plausible null hypothesis, it seems reasonable to expect that "β_j very near 0" is also plausible, and that the distribution of the effects will be unimodal about 0. Alternatively, we can motivate A1 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 assumption A1 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 A1 does not contradict this.

Interestingly, although the unimodal assumption A1 seems natural, much previous related work on estimating False Discovery Rates has assumed, either explicitly or implicitly, that the effect sizes are multimodal (see Results for more details). In contrast, almost all analogous work in large-scale regression assumes a unimodal prior distribution for the effect sizes, common choices being the spike and slab, Laplace, t, normal-gamma, normal-inverse-gamma, or horseshoe priors. One of our aims here is to emphasize potential benefits of also making the unimodal assumption in the FDR context.

Assumption A2 can derived from (**??**) plus a couple of other assumptions. Specifically, assume that a) $\hat{\beta}_j, s_j$ contain most of the information about β_j in the data D_j (i.e. they are "approximately sufficient" for D_j) and that b) s_j alone contains little information about β_j. The former (a) suggests the approximation $\hat{L}_j(\beta_j) \propto p(\hat{\beta}_j, s_j|\beta) = p(\hat{\beta}|\beta, s_j)p(s_j|\beta)$, while the latter (b) suggests assuming that $p(s_j|\beta_j)$ does not depend on β_j. Putting these two together yields $\hat{L}_j(\beta_j) \propto p(\hat{\beta}_j|\beta, s_j)$, which, with (1), implies A2. If $\hat{\beta}_j, s_j$ are not sufficient for D_j then there is some loss of efficiency in making this approximation, but all Bayesian inferences remain valid provided they are interpreted as conditional on $\hat{\beta}_j, s_j$, rather than conditional on D_j. See [**?**] and [**?**] for more general discussion and examples of Bayesian inference based on summary statistics.

Alternatively A2 could be considered definitional for $\hat{\beta}_j, s_j$: that is, choose $\hat{\beta}_j, s_j$ such that A2 holds.

While Assumption A2 has parallels with approaches like [**?**, **?**] that model the Z scores $Z_j = \hat{\beta}_j/s_j$ as normally distributed under the null, there is an important difference: here, observations with a large s_j have an essentially flat likelihood (3), and so do not affect inference for other parameters, whereas when modeling the Z scores directly, observations with large s_j produce $N(0,1)$ Z scores that do affect inference. In this way the likelihood-based approach can take better account of the informativeness of each measurement, as illustrated in Results below.

Implementing the Unimodal assumption

In practice, we implement the unimodal assumption (A1) by assuming a parametric finite mixture for g,

$$g(\cdot;\pi) = \pi_0\delta_0(\cdot) + \sum_{k=1}^{K} \pi_k f_k(\cdot) \tag{4}$$

where $\delta_0(\cdot)$ denotes a point mass at 0, $\pi = (\pi_0, \dots, \pi_K)$ are mixture proportions to be estimated (with $\pi_j \geq 0, \sum_j \pi_j = 1$), and 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 $N(\cdot; 0, \sigma^2)$ denotes the density of a normal distribution with mean 0 and variance σ^2 and $U[\cdot; a, b]$ denotes the density of a uniform distribution on $[a, b]$. That is, the non-zero effects are modeled as either i) a mixture of zero-centered normal distributions; ii) a mixture of zero-centered uniform distributions; or iii) a mixture of "zero-anchored" uniform distributions, which have one end of their range at 0. This approach closely mirrors approaches in [**?**].

To see the flexibility of the representation (4) think of K as being "large", with the values of σ_k^2 or a_k spanning a wide range of values, from essentially zero to unnecessarily large. In this way, as K increases, the three different choices for f_k 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 such as the spike and slab prior, the horseshoe prior **??**, or the more general three-parameter beta prior **??**.

ii) any symmetric unimodal distribution about 0.

iii) any (asymmetric) unimodal distribution about 0.

The latter two claims are related to characterizations of unimodal distributions due to Khintchine (1938) and Shepp (1962); see Feller, 1971, page 158. Furthermore, although it is natural for statisticians to immediately worry about "overfitting" if K is large, in fact the unimodal constraint is sufficient to prevent serious problems with overfitting. For example, note that the unimodal constraint is sufficiently strong that there will exist a non-parametric maximum likelihood estimate (NPMLE) for g under this constraint. (Indeed, if the β_j are observed without error, corresponding to the standard errors $s_j \to 0$, this NPMLE has an explicit and elegant form; [**?**].) Thus allowing K to be large in ii) or iii) can be thought of as approximating the NPMLE for g, under the constraints of symmetry and unimodality (ii), or unimodality alone (iii).

Having emphasized the flexibility that can be achieved with large K, we note that in practice even quite small values of K (e.g. 2-10) provide highly flexible and effective methods. See Section **??** for our software default choices of K and σ_k^2, a_k.

Likelihood for π

Assuming independence across units, the likelihood for π is

$$L(\pi) := p(D_1, \dots, D_J|\pi) = \prod_{j=1}^n p(D_j|\pi) = \prod_j \int p(D_j|\pi, \beta_j) p(\beta_j|\pi) d\beta_j = \prod_j \int p(D_j|\beta_j) g(\beta_j; \pi) d\beta_j \quad (5)$$

Further, the assumptions $\hat{\beta}_j|\beta_j, s_j \sim N(\beta_j, s_j)$ and $\beta_j|s_j, \pi \sim g(\cdot; \pi)$ (given by 4) yield:

$$p(\hat{\beta}_j, s_j|\pi) = \pi_0 N(\hat{\beta}_j; 0, s_j^2) + \sum_k \pi_k \tilde{f}_{kj}(\hat{\beta}_j) \quad (6)$$

where $\tilde{f}_{kj}$ denotes the density of a convolution of f_k with an $N(0, s_j^2)$ density. For example, if $f_k(\cdot) = N(\cdot; 0, \sigma_k^2)$ then the convolution $\tilde{f}_{kj}$ is also normal, $\tilde{f}_{kj}(\cdot) = N(\cdot; 0, \sigma_k^2 + s_j^2)$. If $f_k(\cdot) = U[\cdot; a_k, b_k]$ then $\tilde{f}_{kj}$ is the convolution of a normal with a uniform, which has density

$$\tilde{f}_{kj}(x) = (\Phi(b_k/s_j) - \Phi(a_k/s_j))/(b - a), \quad (7)$$

where Φ is the cumulative distribution function of the standard normal $N(0,1)$ distribution.

Putting these together, the joint model for the unobserved β and the observed $\hat{\beta}, s$ is:

$$p(\hat{\beta}, s, \beta|\pi) = \prod_j g(\beta_j; \pi) p(\hat{\beta}_j, s_j|\beta_j) \tag{8}$$

$$= \prod_j g(\beta_j; \pi) L(\beta_j; \hat{\beta}_j, s_j). \tag{9}$$

We fit this hierarchical model using the following "Empirical Bayes" approach. First we estimate the hyper-parameters π by maximizing the likelihood

$$L(\pi; \hat{\beta}, s) := p(\hat{\beta}, s|\pi) = \int p(\hat{\beta}, s, \beta|\pi) d\beta. \tag{10}$$

This can be done very easily using an EM algorithm. Then, given this estimate $\hat{\pi}$, we compute the conditional distributions

$$p(\beta_j|\hat{\pi}, \hat{\beta}, s) \propto g(\beta_j; \pi) L(\beta_j; \hat{\beta}_j, s_j). \tag{11}$$

In principle we would prefer to take a full Bayes approach that accounts for uncertainty in g (by accounting for uncertainty in π); however, we believe that in most practical applications uncertainty in g will not be the most important concern, and so we compromise this principle for the simplicity of the EB approach. [Note: a Variational Bayes approach to estimating π is also implemented in our R package]

The conditional distributions $p(\beta_j|\hat{\pi}, \hat{\beta}, s)$ encapsulate uncertainty in the values for β_j, combining information across $j = 1, \dots, J$. The combining of the information occurs through estimation of π, which involves all of the data. These conditional distributions can be conveniently summarized in various ways, including point estimates (e.g. the posterior means or medians), and credible intervals/regions.

Efron (2008) 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.

By modeling the z scores directly under the alternative, rather than the z scores as being the truth plus noise, maybe that is what is most problematic? Because the resulting distribution of z scores is just not possible.

Note that [?] 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?

In practice, we currently fix the number of components K to be large, and take the variances $\sigma_1 < \sigma_2 < \dots < \sigma_K$ to be fixed, and vary from very small to very large – sufficiently large that typically $\hat{\pi}_K \approx 0$.

Using a mixture of normal distributions for g has the advantage that, when combined with the normal likelihood, they give an analytic form for the conditional distribution $p(\beta_j|\hat{\pi}, \hat{\beta}_j, s_j)$ (also a mixture of normals).

The local False Sign Rate

The local False Discovery Rate (lfdr) for observation j is

$$\text{lfdr}_j = p(\beta_j = 0|\hat{\beta}, s). \tag{12}$$

The lfdr terminology comes from using "discovery" to refer to rejecting the null ($H_j : \beta_j = 0$). Specifically lfdr_j is the probability that, if we reject H_j, it is a "false discovery".

As pointed out by [?], there are settings where $\beta_j = 0$ is implausible, in which case the lfdr is not useful: if every β_j is non-zero then there is no such thing as a false discovery and the lfdr will always be identically 0. Gelman et al suggest that in such settings we might 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, it could be argued that identifying which are positive and which negative is fundamentally more interesting and useful than identifying which are non-zero. For example, when identifying differentially expressed genes, analysts will often separate the genes that are "upregulated" in one condition from those that "downregulated" in that condition.

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

$$\text{lfsr}_j := \min[p(\beta_j \geq 0|\hat{\beta}, s), p(\beta_j \leq 0|\hat{\beta}, s)]. \tag{13}$$

Thus lfsr_j gives the probability that we would get the sign of β_j wrong if we were to make our best guess. (Note that we count it as 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. We note that the idea of focussing on tail areas, rather than point null hypotheses, has a long history (e.g. Altham? others?).

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.

0.1 Computation Outline

As outlined above, we fit the model using the following Empirical Bayes procedure: 1. Estimate π by maximizing the likelihood $L(\pi)$. 2. Compute the conditional distributions $p(\beta_j|\hat{\beta}_j, s_j, \hat{\pi})$.

Using a normal likelihood $L(\beta_j)$, and assuming g to be a mixture of normal distributions with fixed variances, yields a simple EM algorithm for estimating π in Step 1, and simple analytic forms for the conditional distributions in Step 2.

Results

Improved conservative estimation of π_0

To illustrate estimation of π_0 we provide simulation results for two scenarios:

Scenario 1:

$$f_1(\cdot) = 0.4N(0, 0.25^2) + 0.2N(0, 0.5^2) + 0.2N(0, 1^2), 0.2N(0, 2^2) \tag{14}$$

Scenario 2:

$$f_1(\cdot) = N(0, 4^2) \tag{15}$$

Scenario 1 represents a "difficult" case where many non-zero β are too close to zero to be reliably detected, making reliable estimation of π_0 essentially impossible; Scenario 2 represents an "easier" case where most non-zero β are sufficiently different from zero that they can be reliably detected. making estimation of π_0 easier. For Scenario 1 we considered datasets of size $n = 1000$ (Scenario 1a) and $n = 10,000$ (Scenario 1b); for Scenario 2 we used $n = 1000$.

For each simulation scenario we simulated 200 independent data sets. For each data set we simulated the true value of $\pi_0 \sim U[0, 1]$, and then simulated $\beta_j \sim \pi_0\delta_0 + (1 - \pi_0)f_1(\cdot)$ and $\hat{\beta}_j|\beta_j = N(\beta_j, 1)$ for $j = 1, \ldots, n$. Thus these simulations assume the same precision for each measurement. We applied the methods implemented in the R packages qvalue, fdrtool, locfdr, and mixfdr for estimating π_0. We also applied our method to estimate π_0 in two ways, one using a penalty encourage π_0 to be as large as possible ("null-biased"), and the other using the maximum likelihood estimate for π.

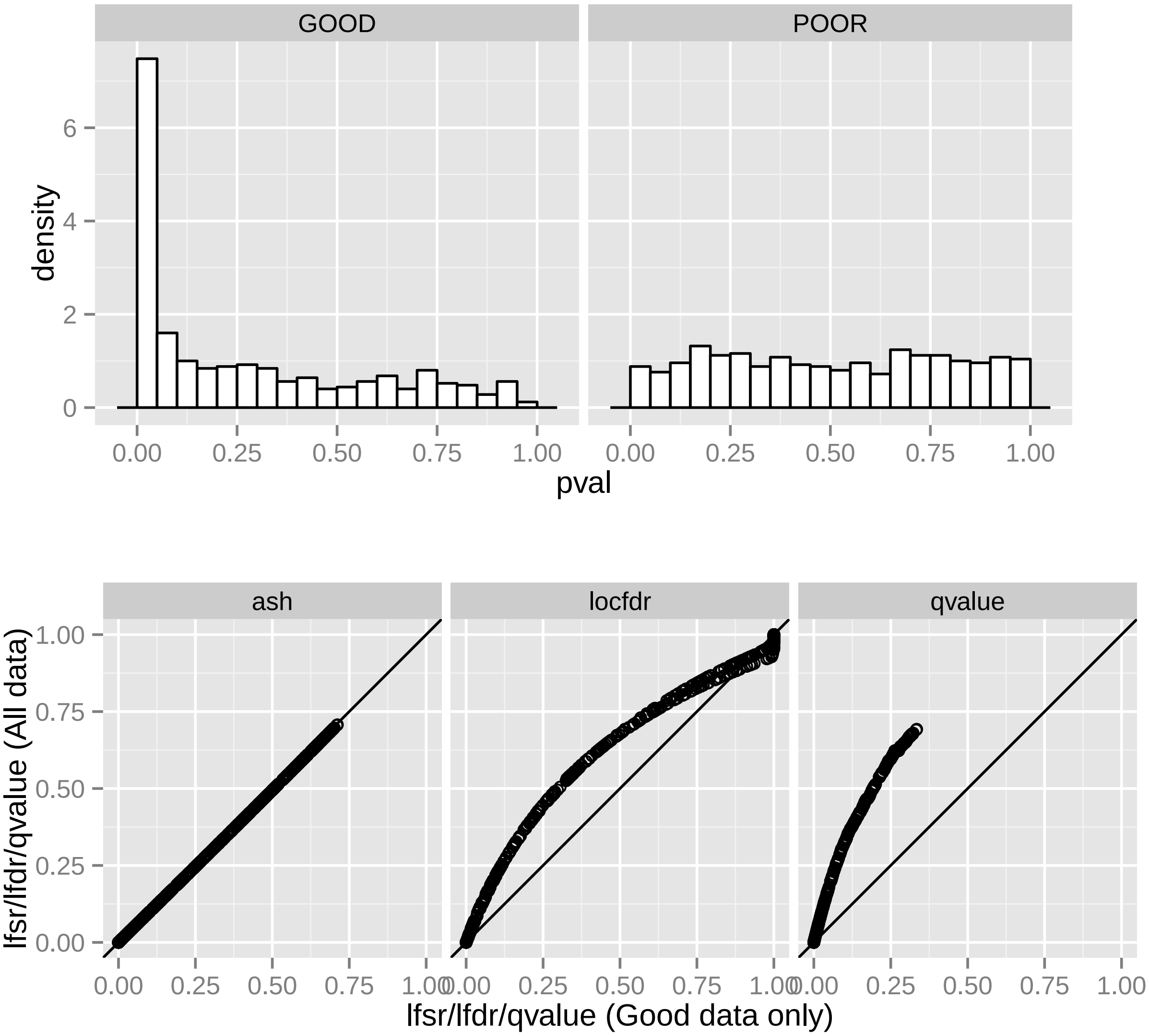

Figure 1

Figure **??** compares estimated and true values of π_0 under each Scenario. For Scenario 1, no method reliably estimates π_0. This is expected since, the scenario was designed to make accurate estimation of π_0 impossible. However, we see that all the methods except for ash (mle) are able to provide a conservative estimate for π_0. Further, among these the estimates from ash (null-biased) are least conservative. Scenario 2 produces similar patterns, except that for this scenario the ash estimates of π_0 are quite accurate (made possible by the fact that most non-zero effects are very different from zero).

These simulations illustrate two key points: i) although π_0 is not identifiable, the penalized likelihood approach provides conservative estimates for π_0; ii) replacing the "zero assumption" with the unimodal assumption can produce less conservative, more accurate, estimates of π_0 (provided of course that the unimodal assumption holds, as it does in these simulations).

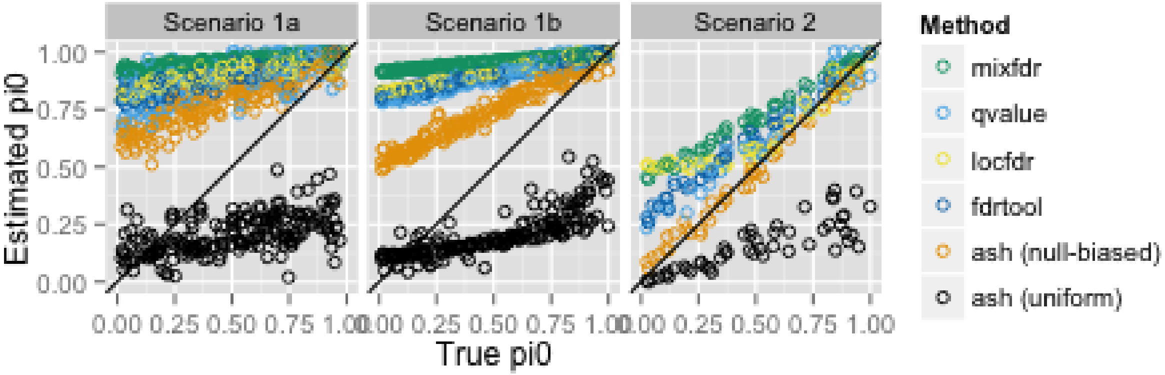

Figure 2. Comparison of true and estimated values of π_0 for simulation scenarios. Scenarios 1a and 1b represent "difficult" scenarios where π_0 is impossible to estimate accurately. However, all methods except the ash mle method are successful in providing conservative estimates for π_0. The ash (null-biased) method is least conservative, and hence most accurate, due to its additional assumption that the effect size distribution g is unimodal.

Note that, even though the exact value of the point mass π_0 cannot, in general, be reliably estimated, the actual underlying distribution g can be quite accurately estimated from the data, provided we assess accuracy by a metric that is not sensitive to π_0: for example by measuring the difference between the true and estimated cumulative distribution function (cdf). See supplementary Figure **??** for examples.

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 FILL IN THESE DETAILS?).

- Illustration that estimated lfsr is more robust.

- Illustration that adjusted estimate of lfsr is still more robust..

- Compare estimated lfsr with true lfsr for difficult simulation case, to show nearly achieves Bayes Risk?

0.2 Accounting for measurement precision improves estimates of FDR/fsr

- Illustration of effects of contamination on FDR by low-quality data comparison with ash.

0.3 Comparison of different underlying component distributions

We simulated data under three scenarios for the underlying effect size distribution g:

Scenario A: A mixture of zero-centered normals,

$$g(\cdot) = (2/3)N(\cdot; 0, 1) + (1/3)N(\cdot; 0, 2^2) \tag{16}$$

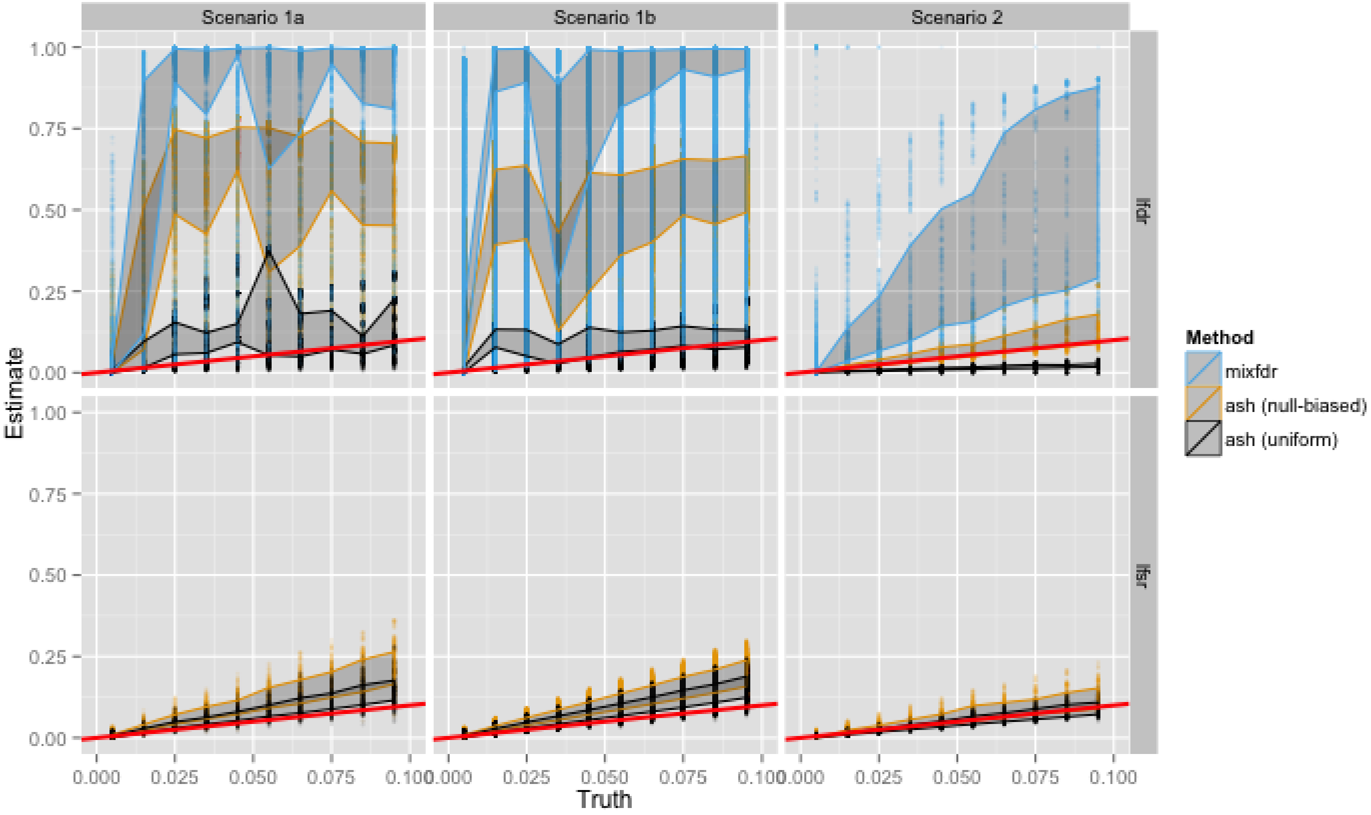

Figure 3. Figure showing lfsr is more robust than lfdr

Scenario B: A symmetric "flat-topped" distribution, designed to be difficult to capture by a mixture of zero-centered normals

$$g(\cdot) = (1/7)[N(\cdot; -1.5, 0.5^2) + N(\cdot; -1, 0.5^2) + N(\cdot; 0.5, 0.5^2) + N(\cdot; 0, 0.5^2) + N(\cdot; 0.5, 0.5^2) + N(\cdot; 1, 0.5^2) + N(\cdot; 1.5, 0.5^2) \tag{17}$$

Scenario C: An asymmetric distribution,

$$g(\cdot) = (1/4)N(\cdot; -2, 2^2) + (1/4)N(\cdot; -1, 1.5^2) + (1/3)N(\cdot; 0, 1) + (1/6)N(\cdot; 1, 1) \tag{18}$$

In each case we simulated $\hat{\beta}_j = \beta_j + N(0, s_j^2 = 1)$.

Because the mixture of zero-anchored uniforms is more flexible than the mixture of zero-centered uniforms, which is more flexible than the mixture of zero-centered normals, we expect that the likelihoods will satisfy $L_{hu} > L_u > L_n$. However, for Scenario A we expect the differences in likelihood to be small, because the true model here is a mixture of zero-centered normals, so any gain of L_{hu} and L_u over L_n is necessarily due to "overfitting". For Scenario B, we expect that L_{hu} and L_u to be significantly larger than L_n, but any increase of L_{hu} over L_u must be due to overfitting (because the truth is symmetric). Finally, for Scenario C we expect L_{hu} to be significantly larger than the others.

Figure **??** illustrates this.

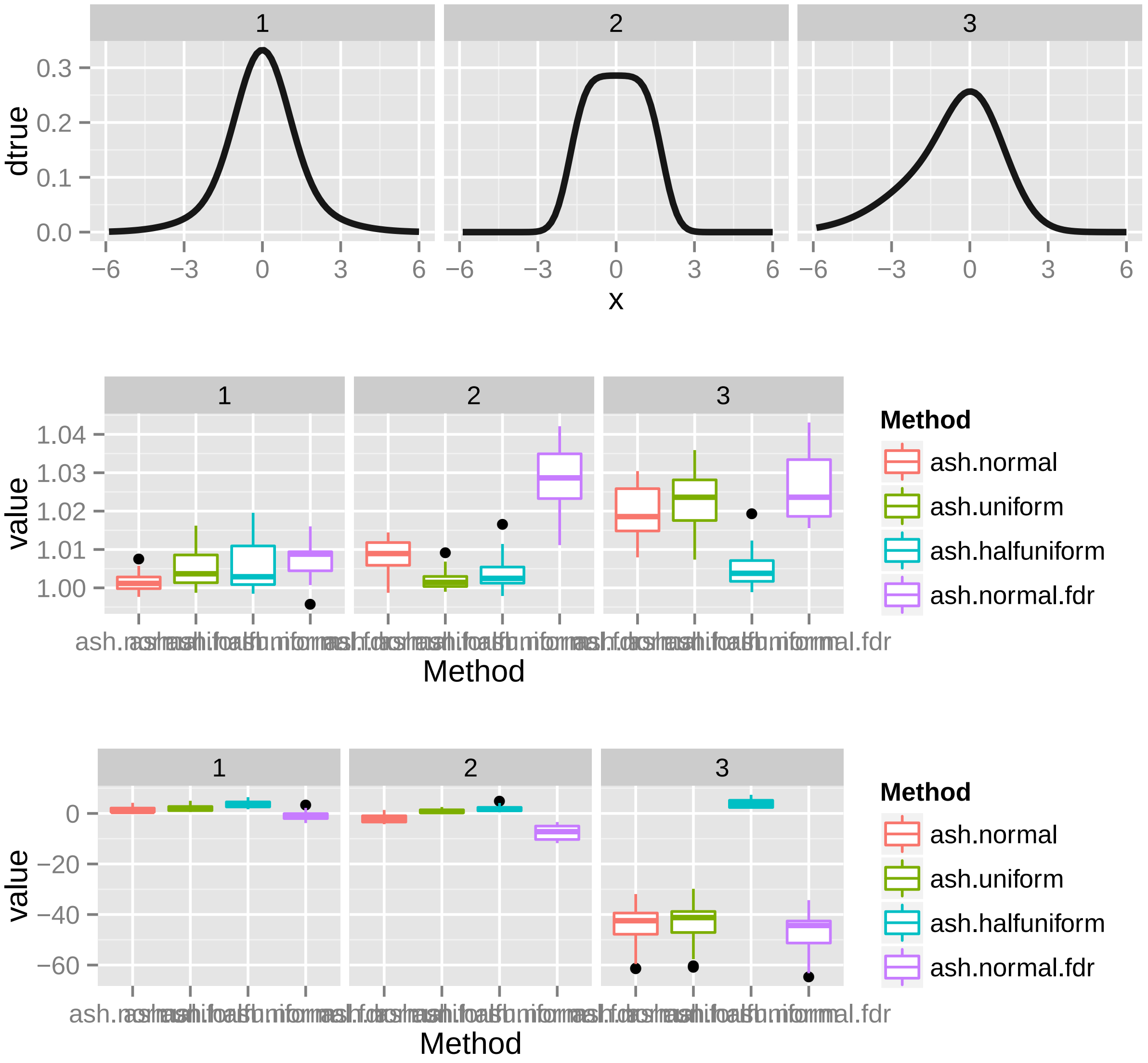

Figure 4

- how does use of the VB approach to estimate pi affect RMSE?

- Comparison between exchangeable effects vs Exchangeable standardized effects; Hedenberg data.

- Possibly difference between CIs obtained by ash and usual CIs (e.g. “of intervals where local fsr is ¡0.05, what proportion are the sign correct”?).

- Asymmetric case (Half uniforms)

Things to emphasise for paper: - the number of components is not critical; the unimodal constraint is enough. You can increase the number of components arbitrarily and the likelihood remains bounded. - the ability to incorporate item-specific measurement error - the ability to compare a model in which effects are proportional to error with model in which effects

- bayesian coverage intervals (attempt to) give much stronger guarantess than standard confidence intervals. Eg among the 0.95 intervals excluding 0, less than 5

Some advantages of the unimodal constraint: -more statsitically efficient (if true); show with small sample sizes. - less sensitive to number of components? Can allow number to tend to infinity? - may be less sensitive to local optima? could demonstrate this by looking at convergence more carefully, and comparing results with random starting points.

- it allows us to easily just vary sigma on a grid, and fit pi, which makes allowing different noise levels really easy!

0.4 The number of components is not critical

OR - this section could be more generally about selecting between models (normal, uniforms etc) using log-likelihood.

Because of the uni-modal constraint, the number of mixture components is generally not critical, at least provided it is "sufficiently large". Indeed, as the number of components tends to infinity, the likelihood is bounded above, and even for large numbers of components and small amounts of data the inferred underlying distributions tend not to be too crazy, showing few signs of "overfitting".

However, it is true that the normal distributions tend to yield smoother estimated underlying distributions. Similarly, using the posterior mean for π, rather than the maximum likelihood estimate, tends to lead to somewhat smoother fitted distributions. In addition, with small amounts of data the underlying distributions are inevitably not well determined by the data, and the fits may vary depending on the underlying assumptions made (e.g. number of components or distribution of the components); however, even then shrinkage-based estimates of the betas can be relatively robust.

0.5 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

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

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$betahat, ss$betasd, 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?

0.6 Side notes on Multiple comparisons

Note on multiple comparisons: it isn’t really a ”problem” but an ”opportunity”. This viewpoint also espoused by Greenland and Robins. 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.

Imagine that you are preparing to perform 1 million tests, each based on a Z score that is assumed to be $N(0,1)$ under the null. You first order these tests randomly, and begin by performing the first test, which returns a Z score of 4. At this point you are interrupted by a friend, who asks how the analysis is going. ”It’s early days, but looking promising” you reply. Well, who wouldn’t? If the aim is to find lots of significant differences, a strong first result is surely a good outcome.

At this point you have reason to expect that many of the subsequent tests also output strong results.

Now consider two alternative scenarios for the remaining 999,999 tests. In the first scenario, the remaining tests produce Z values that fit very well with the null, closely following a standard normal distribution; in the second scenario a large proportion of the remaining tests, say 50 percent, show outcomes that lie outside of $[-4,4]$.

If your friend enquired after your analysis again, your response would surely differ in the first scenario (”Oh, it didn’t pan out so well after all”) vs the second (”It went great”). Further, in the first scenario, if your friend pressed you further about the results of the first test, you would likely, I think, be inclined to put them down to chance. In contrast, in the second scenario, the first test turned out to be, as you hoped, a harbinger of good things to come, and in this scenario you would likely regard that test as likely corresponding to a true discovery.

The key point is that it is the *outcomes* of the tests, not the *number* of tests, that impacts interpretation of that first test.

(Some may by pondering whether the fact that you are about to perform another 999,999 tests should be considered pertinent in responding to your friend. Our view is that it is irrelevant. The standard frequentist framework would disagree, because it requires the analyst to consider hypothetical repetitions of the ”experiment”, and so the fact that the experiment consists of a million tests is pertinent. However, this argument is a distraction from the main point.)

Indeed, we believe that the practice of focussing on the *number* of tests performed is

Focussing on the number of tests performed can be seen as an approximation.

The standard argument is that, when performing multiple tests, some will be significant just by chance.

Acknowledgments

Statistical analyses were conducted in the R programming language [?], Figures produced using the ggplot2 package [?], and text prepared using LaTeX.

Figure Legends

Tables

Supporting Information Legends

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