Empirical Bayes estimation of normal means, accounting for uncertainty in estimated standard errors

Mengyin Lu and Matthew Stephens

e-mail: mengyin@uchicago.edu
mstephens@uchicago.edu

Abstract: We consider Empirical Bayes (EB) estimation in the normal means problem, when the standard deviations of the observations are not known precisely, but estimated with error – which is almost always the case in practical applications. In classical statistics accounting for estimated standard errors usually involves replacing a normal distribution with a $t$ distribution. This suggests approaching this problem by replacing the normal assumption with a $t$ assumption, leading to an “EB $t$-means problem”. Here we show that an approach along these lines can indeed work, but only with some care. Indeed, a naive application of this idea is flawed, and can perform poorly. We suggest how this flaw can be remedied by a two-stage procedure, which first performs EB shrinkage estimation of the standard errors and then solves an EB $t$-means problem. We give numerical results illustrating the effectiveness of this remedy.

Keywords and phrases: Empirical Bayes, shrinkage, normal means, $t$ distributions.

1. Introduction

We consider Empirical Bayes (EB) estimation in applications where we have observed estimates $\hat{\beta}_j$ ($j = 1, \ldots, p$) of a series of underlying “effects” $\beta_j$, with estimated standard errors $\hat{s}_j$. Our goal is to perform EB estimation for the effects $\beta_1, \ldots, \beta_p$, from the observations $\hat{\beta}_1, \ldots, \hat{\beta}_p$, under standard normal theory assumptions, but taking account of uncertainty in the standard errors.

If the standard errors of $\hat{\beta}_1, \ldots, \hat{\beta}_p$ were known, rather than estimated, our problem would simply involve EB inference for the well-studied “Normal means” problem (e.g. Johnstone and Silverman, 2004):

$$\hat{\beta}_j | \beta_j, s_j \sim N(\beta_j, s_j^2), \quad j = 1, \ldots, p; \tag{1}$$
$$\beta_j | s_j \sim g_\beta \in \mathcal{G}, \quad j = 1, \ldots, p; \tag{2}$$

where $s_j$ denotes the “true” standard error of $\hat{\beta}_j$, and $\mathcal{G}$ is some specified family of distributions. (The conditioning on $s_j$ in (2) makes explicit an assumption that the $\beta_j$ are independent and identically distributed from $g_\beta$, independent of $s_j$, an assumption we relax later.) Fitting this EB model involves first obtaining an estimate $\hat{g}_\beta$ for $g_\beta$ (e.g. by marginal maximum likelihood), and then computing the posterior distributions $p(\beta_j | \hat{\beta}_j, s_j, \hat{g}_\beta)$. These posterior distributions can be used to obtain both point and interval estimates of $\beta_j$. And, if the family $\mathcal{G}$ involves sparse distributions (with a point mass on 0), then the posterior distributions can also be used to compute (local) false discovery rates (Efron, 2004), effectively providing an EB solution to the “multiple testing” problem.

*This work was supported by NIH grant HG002585 and by a grant from the Gordon and Betty Moore Foundation.
This EB normal means problem is well studied, and there exist flexible software implementations for solving it for a range of choices of $G$ (e.g. Stephens, 2016). For example, methods in Stephens (2016) effectively solve this problem for $G$ the set of all unimodal distributions (by exploiting the fact that any such distribution can be approximated, to arbitrary accuracy, by a mixture of sufficiently many uniform distributions).

In classical statistics, the fact that standard errors are estimated is usually dealt with by replacing normal distributions with $t$ distributions. Indeed, in the settings we consider here, we have

$$(\hat{\beta}_j - \beta_j)/s_j \sim N(0,1),$$

and

$$(\hat{\beta}_j - \beta_j)/\hat{s}_j \sim t_\nu,$$

where $t_\nu$ denotes the $t$ distribution on $\nu$ degrees of freedom. Expression (4) is routinely used in classical statistics to obtain confidence intervals for $\beta_j$ and $p$ values testing $\beta_j = 0$.

From this it is tempting to replace the EB normal means problem (1)-(2) with what we call the “EB $t$-means problem” (EBTM):

$$\hat{\beta}_j | \beta_j, \hat{s}_j \sim t_\nu(\beta_j, \hat{s}_j)$$

$$\beta_j | \hat{s}_j \sim g_\beta \in G,$$

where $t_\nu(\mu, \sigma)$ denotes the generalized $t$ distribution on $\nu$ degrees of freedom, with mean $\mu$ and scale parameter $\sigma$ (i.e. the distribution of $\mu + \sigma T$ when $T \sim t_\nu$). While this EBTM problem is much less studied than the EB normal means problem, Stephens (2016) also provides flexible software implementations solving the EBTM problem – estimating $g$ and computing posterior distributions $p(\beta_j | \hat{\beta}_j, \hat{s}_j, \hat{g})$ – for a range of choices of $G$.

Unfortunately, there is a problem with this tempting naive approach: while the EBTM problem (5)-(6) is well-defined and solvable, the standard theory that leads to (4) does not imply (5). The reason is that in (4) $\hat{s}_j$ is random, and not conditioned on, and the unconditional expression does not imply a corresponding conditional one:

$$(\hat{\beta}_j - \beta_j)/\hat{s}_j \not\Rightarrow (\hat{\beta}_j - \beta_j)/\hat{s}_j | \hat{s}_j \sim t_\nu.$$

To give a simple explicit example of this: if $\hat{\beta}_j \sim N(0,1)$ and $\hat{s}_j^2 \sim \chi^2_1$ then $\hat{\beta}_j/\hat{s}_j \sim t_1$ but $\beta_j/\hat{s}_j | \hat{s}_j \sim N(0,1/\hat{s}_j^2)$. Consequently (5) does not hold in general, and – as we show later – ignoring this can produce very unreliable inferences in practice.

In this paper we describe a simple solution to this problem. Our solution involves EB analysis of the standard errors $\hat{s}_j$ (Smyth, 2004), which is already widely used in genomics applications – indeed, currently much more widely used than EB analysis of the effect estimates $\hat{\beta}_j$. Our approach effectively combines the methods from Smyth (2004) with the methods for the EBTM problem from Stephens (2016). We demonstrate empirically that, in contrast with the naive approach, this combined approach can provide reliable inference.

2. Methods

Assume that, independently for $j = 1, \ldots, p$, we have observed estimates $\hat{\beta}_j$ and corresponding (estimated) standard errors $\hat{s}_j$, satisfying

$$p(\hat{\beta}_j, \hat{s}_j | \beta_j, s_j) = p(\hat{s}_j | s_j)p(\hat{\beta}_j | \beta_j, s_j)$$

$$p(\hat{\beta}_j, \hat{s}_j | \beta_j, s_j) = p(\hat{s}_j | s_j)p(\hat{\beta}_j | \beta_j, s_j)$$

2
where

\[ \hat{\beta}_j | \beta_j, s_j \sim N(\beta_j, s_j^2) \quad (9) \]

\[ \hat{s}_j^2 | s_j \sim s_j^2 \chi^2_p / \nu. \quad (10) \]

For example (8)-(10) hold if \( \hat{\beta}_j, \hat{s}_j \) are the usual estimate of \( \beta_j \) and its standard error in a simple linear regression, \( y_j = x \beta_j + e_j \), where \( y_j \) and \( x \) are observed n-vectors and the residual errors \( e_j \) are normally distributed with mean 0 and variance \( \sigma^2_j \).

Our goal is to perform EB estimation for \( \beta_1, \ldots, \beta_p \) under the assumption (6) that \( \beta_j | s_j \sim g_{\beta} \in G \).

As noted in the Introduction, if (5) held then this would be solved by methods for the EBTM problem in Stephens (2016). However, unfortunately (8)-(10) do not imply (5) and so (5) does not hold in general.

We now describe a simple solution to this problem, based on combining the EBTM methods in Stephens (2016) with EB estimation for \( \hat{s}_j \) using the methods in Smyth (2004). Specifically, Smyth (2004) combines the sampling distribution for \( \hat{s}_j \) (10) with an assumption that the true variances \( s_j^2 \) come from an inverse-gamma distribution, which can be written:

\[ s_j^{-2} \sim s_0^{-2} \chi^2_{\nu_0} \quad (j = 1, \ldots, p) \quad (11) \]

where \( s_0, \nu_0 \) are parameters to be estimated. The EB approach in Smyth (2004) estimates \( s_0, \nu_0 \) from the observations \( \hat{s}_1, \ldots, \hat{s}_p \) (using a method of moments), and then bases inferences for \( s_j^2 \) on its posterior distribution given these estimates, which is also an inverse-gamma distribution. Indeed, given \( s_0 \) and \( \nu_0 \), the posterior can be written

\[ s_j^{-2} | \hat{s}_j \sim \tilde{s}_j^{-2} \chi^2_{\tilde{\nu}_j} / \tilde{\nu}_j, \quad (12) \]

where

\[ \tilde{\nu}_j := \nu_0 + \nu_j \quad (13) \]

\[ \tilde{s}_j^2 := (\nu_0 s_0^2 + \nu_j \hat{s}_j^2) / (\nu_0 + \nu_j). \quad (14) \]

In particular Smyth (2004) uses \( \tilde{s}_j^2 \) – which lies between \( s_0^2 \) and \( \hat{s}_j^2 \) – as a “moderated” estimate of \( s_j^2 \).

The key to our approach is the following simple Lemma.

**Lemma 1.** Assuming (8),(9) and (12) it follows that

\[ \hat{\beta}_j | \beta_j, \tilde{s}_j \sim t_{\tilde{\nu}_j}(\beta_j, \tilde{s}_j). \quad (15) \]

Thus, although (5) does not hold in general, under the assumptions (8)-(11) (which imply (12)) an analogous expression (15) does hold. This analogous expression simply involves replacing the original standard errors and degrees of freedom with their moderated values, (13) and (14). Combining (15) with (6) then yields an EBTM problem that can be solved using methods from Stephens (2016).

**2.1. A two-step strategy**

Putting this all together, we suggest the following two-step strategy for fitting the EBNM model, accounting for uncertainty in estimated standard errors:

1. Estimate the linear regression parameters \( \hat{\beta}_j, \hat{s}_j \).
2. Test the moderated standard errors \( \tilde{s}_j \) obtained from (12).
3. Perform EBTM for the moderated standard errors and degrees of freedom.
4. Use the moderated estimates to update the linear regression parameters.

This two-step strategy provides a practical solution to the problem of accounting for uncertainty in estimated standard errors, while also leveraging the strengths of both EB and EBTM approaches.
1. Apply EB shrinkage methods to estimated standard errors \( \hat{s}_1, \ldots, \hat{s}_p \), using the likelihood (10) and prior (11), as in Smyth (2004). This yields estimates for \( s_0, \nu_0 \), and subsequently moderated estimates \( \tilde{s}_j \) (14) and degrees of freedom \( \tilde{\nu}_j \) (13).

2. Apply methods for the EBTM problem (e.g. Stephens, 2016) to the estimates \( \hat{\beta}_j \), estimated standard errors \( \tilde{s}_j \) and degrees of freedom \( \tilde{\nu}_j \). This yields estimates for \( g_\beta \) and the posterior distributions \( p(\beta_j | \hat{\beta}_j, \tilde{s}_j, g_\beta) \).

**Notes**

1. Like many two-step procedures, this two-step procedure is not fully efficient: in principle it would be more efficient to jointly estimate \( g_\beta, s_0, \nu_0 \) from \( (\hat{\beta}_1, \hat{s}_1), \ldots, (\hat{\beta}_p, \hat{s}_p) \), rather than first estimate \( s_0, \nu_0 \) from \( \hat{s}_1, \ldots, \hat{s}_p \) and then estimate \( g_\beta \) while fixing the estimates of \( s_0, \nu_0 \). However in practice, because the true number of parameters \( p \) is typically large, \( s_0, \nu_0 \) can already be accurately estimated from \( \hat{s}_1, \ldots, \hat{s}_p \), and in our view the convenience of the two-step procedure greatly outweighs any minor loss of efficiency.

2. The distributional assumption (11), which leads to (12), may seem somewhat restrictive. However, the moderated \( t \) statistics from Smyth (2004) – which rely on the same assumption – have been found to be well behaved in practice and are widely used. See Lu and Stephens (2016); Phipson et al. (2016) for discussion and assessment of more flexible assumptions.

3. Although assumptions (11) and (10) are the simplest way to obtain posterior distributions of the form (12), the form (12) holds more generally. For example, the *voom* framework (Law et al., 2014) adapts methods in Smyth (2004) to deal with the count nature of RNA sequencing data, and involves both accounting for mean-variance relationships and using weighted least squares rather than ordinary least squares. However, it ultimately yields conditional distributions of the form (12), which – by Lemma 1 – lead to an EBTM problem for \( \beta_1, \ldots, \beta_p \).

### 2.2. Dependence of \( \beta_j \) on \( \hat{s}_j \)

Equation (6) assumes that the \( \beta_j \) are independent of \( \hat{s}_j \). Methods in Stephens (2016) for the EBTM problem can deal with the more general assumption:

\[
\beta_j / \hat{s}_j^\alpha | \hat{s}_j \sim g_\beta \in \mathcal{G},
\]

for any choice of \( \alpha \in \mathcal{R} \). The choice \( \alpha = 0 \) gives (6). The choice \( \alpha = 1 \) corresponds to assuming that the moderated \( t \) statistics from Smyth (2004) are independent of \( \hat{s}_j \), which in turn leads to the property that EB measures of significance (e.g. local FDR) are monotonic as the moderated \( t \) statistics move away from 0 (and monotonic in the corresponding \( p \) values if \( g_\beta \) is symmetric about 0). Thus \( \alpha = 1 \) can be thought of as corresponding to the implicit assumption made when ranking significance by \( p \) values from the moderated \( t \) statistics (Wakefield, 2009).

Although values of \( \alpha \) other than 0 and 1 do not have a straightforward motivation or interpretation, it is straightforward to fit these models, and to estimate \( \alpha \) by comparing likelihoods if desired.
2.3. An ad hoc strategy that avoids the EBTM problem

The framework outlined above has the advantage of being based on clear statistical principles. However, it has the disadvantage that the EBTM problem is often more complex to solve than the EBNM problem. In our numerical studies below we therefore also consider an alternative ad hoc strategy, which avoids solving the EBTM problem.

This ad hoc strategy starts by using the same ideas as above to obtain the estimates \( \hat{\beta}_j \), moderated standard errors \( \hat{s}_j \) and degrees of freedom \( \hat{\nu}_j \). However, rather than applying the EBTM methods to these data, we convert the problem into a “normal means” problem, by changing the standard errors. Specifically for each \( j \) we define the “adjusted standard error” \( s'_j \) to be the value for which the z-score \( \hat{\beta}_j / s'_j \) results in the same \( p \)-value (when compared with a standard normal distribution) as from the moderated \( t \) test (comparing \( \hat{\beta}_j / \hat{s}_j \) with a \( t \) distribution on \( \hat{\nu}_j \) degrees of freedom).\(^1\) We then use \( \hat{\beta}_j, s'_j \) as the inputs to an EBNM problem to obtain posterior distributions and shrinkage estimates for \( \beta_j \).

3. Numerical Studies

We illustrate our two-stage strategy, and compare it with the naive strategy, the ad hoc strategy, and other related methods, using simulations. To make our simulated standard errors and test statistics realistic, we base our simulations on real data from an RNA sequencing experiment (RNA-seq data). However, unlike real RNA-seq data, our simulations create data that are independent across genes. In practice RNA-seq data are often strongly correlated among genes, and these correlations can cause severe complications for many analyses methods (Leek and Storey, 2007), including the Empirical Bayes methods used here (Efron, 2010; Gerard and Stephens, 2018). By removing these correlations here we are comparing methods under idealized conditions, and seek to show that even under idealized conditions the naive approach – which does not use EB shrinkage of the standard errors – performs poorly. For empirical comparisons of methods on correlated RNA-seq data see Gerard and Stephens (2018); Lu (2018).

We perform simulations for two groups, each containing \( n \) samples, with \( n = 2, 4, 10 \) and \( p = 10,000 \) genes. The effects are simulated from \( \pi_0 \delta_0 + (1 - \pi_0)g_1 \) for various choices of distribution \( g_1 \) (Figure 1; Table 1), and then divided by a scaling factor \( S_n \) chosen so that power is similar for different \( n \) (\( S_2 = 0.125, S_4 = 0.5, S_{10} = 1.5 \)). For each combination of \( (n, g_1) \), we simulate 50 datasets with \( \pi_0 \) drawn uniformly from \([0,1]\).

| 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  | \( 2/3N(0,1^2) + 1/3N(0,2^2) \)                                  |
| flat-top     | \( (1/7)[N(-1.5,5^2) + N(-1.5^2) + N(-.5,5^2) + N(0,5^2) + N(0.5,5^2) + N(1.0,5^2) + N(1.5,5^2)] \) |
| big-normal   | \( N(0,4^2) \)                                                   |
| bimodal      | \( 0.5N(-2,1^2) + 0.5N(2,1^2) \)                                 |

Table 1: Summary of simulation scenarios considered

\(^1\)The following R function computes the adjusted standard error from an effect estimate \( bhat \) and corresponding \( p \) value \( p \): \( \text{pval2se} = \text{function}(bhat,p)\{z = \text{qnorm}(1-p/2) ; s = abs(bhat/z) ; \text{return}(s) \} \).
We analyzed each simulated dataset with several methods based on the voom-limma (VL) pipeline (Law et al., 2014), which uses the voom and lmFit functions from the limma R package (Ritchie et al., 2015) to obtain estimates $\hat{\beta}$ and standard errors $\hat{s}$, along with degrees of freedom $\nu$. Many of the pipelines also use the eBayes function to obtain moderated standard errors $\tilde{s}$ and moderated degrees of freedom $\tilde{\nu}$, which yield moderated $t$ statistics and corresponding $p$ values.

- **VL+ash**: this is the “naive” approach, which directly feeds the $\hat{\beta}$, $\hat{s}$, $\nu$ (without variance moderation) into the EBTM solver in the ash function in the ashr software (Stephens, 2016). As noted above this approach is flawed in principle, and our results show it can also perform poorly in practice.
- **VL+eBayes+ash** and **VL+eBayes+ash.alpha=1**: these are our proposed pipelines, which feed the $\hat{\beta}$ and moderated standard errors $\tilde{s}$ (and $\tilde{\nu}$) to the ash EBTM solver (with $\alpha = 0$ for VL+eBayes+ash and $\alpha = 1$ for VL+eBayes+ash.alpha=1).
- **VL+pval2se+ash**: this is our “ad hoc” approach (Section 2.3), which converts the EBTM problem into an EBNM problem by computing “adjusted standard errors” $s'$, and then applies ash to solve the EBNM problem for $(\hat{\beta}, s')$.
- **VL+eBayes+qvalue**: this is a standard pipeline for controlling FDR in differential expression studies (not based on EB methods): it feeds the $p$ values from the moderated $t$ statistics to the qvalue software (Storey, 2002), which outputs an estimate for $\pi_0$ and a $q$-value for each test which can be used to control FDR.

**Estimation of null proportion**  All of the above methods provide an estimate of the null proportion, $\pi_0$. Obtaining accurate estimates of $\pi_0$ is important for obtaining accurate estimates of FDR: underestimating $\pi_0$ will lead to anti-conservative estimates of FDR, whereas overestimating $\pi_0$ will lead to conservative (over-)estimates of FDR, effectively reducing statistical power.

Figure 2 compares the estimated $\pi_0$ with the true $\pi_0$ in our simulations. The first key observation is that the naive approach voom+ash can dramatically underestimate $\pi_0$, and cannot be recommended. All other approaches generally provide reasonable (conservative) estimates of $\pi_0$, with the ash-based approaches producing more accurate (less conservative) estimates than those from qvalue. This improved accuracy comes from the additional assumption made by the EB approach in ash, that the effects are unimodal (Stephens, 2016). The results are reasonably robust to this assumption, but estimates of $\pi_0$ can be anti-conservative in the bimodal scenario (just as in Stephens (2016)).

**Assessment of FDR control and power**  Figures 3 assesses how well each method controls FDR in our simulations (at nominal level 0.05), and Figure 4 shows the corresponding power (proportion...
of true effects declared significant). The naive method completely fails to control FDR for very small sample sizes (2 vs 2 or 4 vs 4). Other methods perform generally well at controlling FDR, although there is some lack of FDR control of ash-based methods in the bimodal scenario. The ash-based methods are slightly more powerful than qvalue because of the less conservative estimates of π₀.

**Effect estimates** One attractive feature of the EB approach to multiple testing is that it provides not only estimates of FDR, but also shrinkage estimates of effect sizes. To compare the accuracy of the shrinkage estimates with the original (unshrunk VL+eBayes) estimates we compute the relative root mean squared error (RRMSE) for each method as the ratio of the method’s RMSE and the baseline RMSE for the original estimates. Here RMSE := \( \sqrt{\sum_j (\tilde{\beta}_j - \beta_j)^2} \).

The results (Figure 5) demonstrate the expected benefits of shrinkage estimation: the shrunken estimates from VL+eBayes+ash (whether \( \alpha = 0 \) or 1) are consistently better than the original unshrunk estimates. The improvement on the baseline RMSE is up to 90% in settings where most effects are null, where the benefits of shrinkage are strongest.

**Calibration of posterior intervals** In addition to shrinkage point estimates of \( \beta \), the EB approach also provides “shrunk” interval estimates. Stephens (2016) used simulations to show that, under idealized conditions (with known standard errors), these interval estimates not only have good coverage properties on average, but also “post-selection”: that is, even if we focus only on significant effects, the coverage of the EB credible intervals is good. This property is difficult to obtain in other ways.

Here we repeat this coverage assessment in the case of estimated standard errors. Table 2 shows the coverage rates of 95% lower credible bounds for the effects, split into all observations (a), significant negative discoveries (b) and significant positive discoveries (c). Note that for significant negative discoveries (b), the lower credible bound is bounding how “large” the effect is (in absolute value), whereas for positive discoveries (c) it is bounding how close to 0 it can be. In general coverage rates are satisfactory, with the most prominent exception being the case \( N = 2 \) in (b), where coverage rates are often much lower than the nominal 95%. This says that the method is “over-shrinking” the significant effects towards zero in this case, probably due to underestimating the length of the tail of the effects. In low-signal situations some level of over-shrinkage may be inevitable if we want to maintain conservative behaviour (i.e avoid under-shrinkage); thus it is unclear to what extent this behavior could be improved on.

### 3.1. Discussion

In summary, we have shown how EB analysis of normal means with estimated standard errors can be satisfactorily solved by performing an EB analysis of the “t-means” problem (EBTM), but only after applying EB methods to the estimated standard errors themselves to obtain moderated estimates of the standard errors (and associated degrees of freedom).

Our numerical results also show that a simpler ad hoc approach, VL+pval2se+ash in the Figures, which avoids solving the more complex EBTM problem by instead adjusting the standard errors and solving an EBNM problem, can work adequately to control false discovery rates. However, it performs less well in estimation accuracy than the more principled approaches based on solving the EBTM problem.

Code used to obtain the numerical results presented here is available at [http://doi.org/10.5281/zenodo.2547022](http://doi.org/10.5281/zenodo.2547022).
Figure 2: Comparison of true and estimated values of π₀ on simulated data. Generally VL+ash is anti-conservative, often substantially under-estimating π₀. When the UA holds the other three methods yield conservative (over-)estimates for π₀, with VL+eBayes+ash, VL+eBayes+ash.alpha=1 and VL+pval2se+ash being less conservative, and hence more accurate. When the UA does not hold ("bimodal" scenario) the VL+eBayes+ash estimates are slightly anti-conservative.
Figure 3: Comparison of empirical false discovery proportions (FDP), at $q$-value < 0.05, on simulated data. Generally the naive method, $VL+ash$, is anti-conservative, failing to control FDP < 0.05. In contrast, other methods generally control FDP near or under 0.05, although $VL+eBayes+ash$ is slightly anti-conservative in the “big-normal” scenario with small sample size (2 vs 2).
Figure 4: Comparison of proportion of discoveries ("power"), at $q$-value $< 0.05$, on simulated data. Typically $VL+eBayes+ash$ and $VL+eBayes+ash.alpha=1$ have more discoveries than $VL+eBayes+qval$, while controlling FDP (Figure 3).
Figure 5: Comparison of RRMSE (relative root mean squared error) of effect estimates on simulated data. To compute RRMSE we compute RMSE of VL as the baseline level, and divide the RMSE of each method by this baseline. Thus by definition RRMSE of VL is 1. VL+EBayes+ash is more accurate (RRMSE<1) in all scenarios, especially when π₀ is close to 1. The ad hoc approach VL+pval2se+ash performs less well for small sample sizes (although similar to VL+EBayes+ash for 10 vs 10, except for “big-normal” scenario).
(a) All observations. Coverage rates are generally satisfactory, except for the big-normal scenario case when \( N=2 \).

|           | big-normal | bimodal | flat-top | near-normal | spiky |
|-----------|------------|---------|----------|-------------|-------|
| \( N=2 \) | 0.80       | 0.93    | 0.96     | 0.91        | 0.93  |
| \( N=4 \) | 0.93       | 0.96    | 0.96     | 0.96        | 0.95  |
| \( N=10 \)| 0.97       | 0.97    | 0.96     | 0.96        | 0.95  |

(b) “Significant” negative discoveries. Coverage rates are generally satisfactory when \( N = 10 \) and for \( N = 4 \) (except for big-normal scenario), but are generally poor for \( N = 2 \), suggesting over-shrinkage (underestimating the length of the tail of \( q \)) in this setting.

|           | big-normal | bimodal | flat-top | near-normal | spiky |
|-----------|------------|---------|----------|-------------|-------|
| \( N=2 \) | 0.23       | 0.74    | 0.94     | 0.65        | 0.74  |
| \( N=4 \) | 0.76       | 0.95    | 0.94     | 0.95        | 0.94  |
| \( N=10 \)| 0.95       | 0.94    | 0.94     | 0.94        | 0.93  |

(c) “Significant” positive discoveries. Coverage rates are generally satisfactory.

Table 2: Table of empirical coverage for nominal 95% lower credible bounds for method \( VL+eBayes+ash \) on simulated data.
References

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* 99(465), 96–104.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association* 105(491), 1042–1055.

Gerard, D. and M. Stephens (2018). Empirical Bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *Biostatistics*, kxy029.

Johnstone, I. M. and B. W. Silverman (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics* 32(4), 1594–1649.

Law, C. W., Y. Chen, W. Shi, and G. K. Smyth (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15(2), R29.

Leek, J. T. and J. D. Storey (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3(9), e161.

Lonsdale, J., J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, et al. (2013). The genotype-tissue expression (GTEx) project. *Nature genetics* 45(6), 580–585.

Lu, M. (2018). *Generalized Adaptive Shrinkage Methods and Applications in Genomics Studies*. Ph. D. thesis, University of Chicago.

Lu, M. and M. Stephens (2016). Variance adaptive shrinkage (vash): flexible empirical Bayes estimation of variances. *Bioinformatics*, btw483.

Phipson, B., S. Lee, I. J. Majewski, W. S. Alexander, and G. K. Smyth (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The annals of applied statistics* 10(2), 946.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research* 43(7), e47.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3(1), Article 3.

Stephens, M. (2016). False discovery rates: a new deal. *Biostatistics* 18(2), 275–294.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3), 479–498.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genetic epidemiology* 33(1), 79–86.
Appendix A: Simulation details

The following simulation scheme is designed to create realistic count datasets that mimic the structure of real RNA-seq data, making distributional assumptions only sparingly.

1. Start with real RNA-seq data on \( J \) genes in \( N \) samples.
2. Create a “null” data set containing two groups (\( A \) and \( B \)), of sizes \( n_A, n_B \), by randomly sampling (without replacement) \( n_A \) samples for group \( A \) and \( n_B \) samples for group \( B \). Because the assignment of samples to the two groups is random, this is a null dataset by construction. Let \( C_{ji} \) denote the read count for gene \( j \) and sample \( i \).
3. Randomly select \( J(1-\pi_0) \) genes as “alternative genes”, and generate their effects (log\(_2\)-fold-change between groups) \( \beta_j \)'s from a specified “effect distribution” \( g_1 \).
4. For these alternative genes, if \( \beta_j > 0 \) (so group \( B \) should be more highly expressed), we use Poisson thinning to achieve the desired fold-change \( 2^{\beta_j} \), i.e. thin the read counts in group \( A \) as follows:
   \[
   C_{ji}^* \sim \text{Binomial}(C_{ji}, 2^{-\beta_j}), \quad \forall i \in A.
   \] (17)
   Similarly if \( \beta_j < 0 \), thin the read counts in group \( B \):
   \[
   C_{ji}^* \sim \text{Binomial}(C_{ji}, 2^{-\beta_j}), \quad \forall i \in B.
   \] (18)

Replacing \( C_{ji} \) by \( C_{ji}^* \) will result in a new RNA-seq dataset, where the true effects follow \( \pi_0 \delta_0 + (1-\pi_0)g_1 \).

Simulations

The above simulation scheme, which we developed during our work on this project, was used by Gerard and Stephens (2018) to generate realistic simulated RNA-seq datasets with a desired effect distributions, while still preserving most of the structure (correlation, magnitude, etc) of the actual RNA-seq data. Unfortunately correlations among genes create substantial complications for many analysis methods (Leek and Storey, 2007), including ours; see Gerard and Stephens (2018) for extensive discussion and further references. To avoid these complications here we modify this scheme to remove correlations between genes. Specifically we modify step 2 to randomly select the \( n_A \) and \( n_B \) samples for groups \( A \) and \( B \) independently at each gene. This modification ensures that the simulated null data at each gene are independent.

While this modification makes the simulations unrepresentative of typical RNA-seq experiments (since real data are typically correlated across genes), it allows us to study the behaviour of methods under idealized situations, which is helpful for understanding the main conceptual contribution of our work here. Results of our methods on the more realistic simulations with correlations intact are given in Lu (2018).

Our simulations here used RNA-seq data from liver tissue samples distributed by the Genotype-Tissue Expression (GTEx) project (Lonsdale et al., 2013). These data (GTEx V6 dbGaP accession phs000424.v6.p1, release date: Oct 19, 2015, http://www.gtexportal.org/home/) contained data on 119 samples, and we restricted simulations to the 10,000 top expressed genes.