Master's Paper of the Department of Statistics, the University of Chicago
(Internal document only, not for circulation)

Degree Papers for Masters in Statistics

— EbayesThresh with Heterogeneous Variance

Kan Xu

Advisor: Matthew Stephens

Approved _____

Date _____

April-9, 2017

Abstract

This paper is an extension of EbayesThresh (the empirical bayesian thresholding method) under assumption of data with homogeneous standard deviation noises described in Johnstone and Silverman (2004)[1]. We ease this restriction, allowing heterogeneous standard deviation of noises, and provide details in estimating the true effects with posterior estimator, e.g. posterior mean or posterior median. The performance of the model with heterogeneous standard deviation is compared to that of the model assuming homogeneous standard deviation. We also discuss the implication of threshold constraints on estimation of non-null weight and the effect on estimation efficiency in terms of mean squared error and mean absolute error in the article.

Contents

1 Introduction

1.1 Background

In many practical settings, the true effect of an event is sparse in the sense that the effect on most of the subjects are zero (or very near to zero) with observations of the event subject to noise. A common example would be in the astronomical context with noisy observations of the pixels of an image of which only a few are expected to be observed from true signals[1]. A simple model to depict such cases is

$$X_i = \mu_i + \epsilon_i, \ \ i = 1, 2, \cdots, n \tag{1.1}$$

in which X_i is the observation of subject i, μ_i is the true effect of the event, and ϵ_i is the noise. If the noises are i.i.d. from a mean 0 and variance 1 distribution, estimating μ_i simply with x_i will produce a mean square error of 1. However, the fact of sparsity, which can be inferred from the observations, provides space for building a model subsuming this extra information. One potential technique is to filter out null signals with threshold[1]. For example, given threshold t_i for each i, any observation X_i satisfying $|X_i| < t_i$ would be treated as null signal. The difficult part is to choose a proper threshold.

In Johnstone and Silverman (2004)[1], the authors discuss posterior median thresholding on data with noises of homogeneous variance. In practice, it is common to have observations with noises from heterogeneous sources. In this paper, we expand this thresholding method to the heterogeneous variance case.

1.2 Empirical Bayesian Method

Assume errors are independent Gaussian white noise with heterogeneous variance

$$X_j | \mu_j, s_j \sim N(\mu_j, s_j^2). \tag{1.2}$$

The true effect on each subject j is μ_j. The idea of many null effects is reflected in the model from a Bayesian perspective that the true effect μ_j has a prior of a mixture distribution, with a probability mass at zero. A heavy-tailed unimodal symmetric density conditional on non-zero effect is assigned to capture potential outliers. Following Johnstone and Silverman (2004)[1], we use Laplace distribution

$$\mu_j \sim (1 - w)\delta_0(\mu_j) + w\gamma_a(\mu_j) \tag{1.3}$$

in which

$$\gamma_a(x) = \frac{1}{2} a e^{-a|x|} \tag{1.4}$$

is the Laplace density with zero mean and parameter a and $1-w$ is the weight of mass at zero.

The objective is to compute posterior estimator of μ_j given information observed, X_j and s_j, using its posterior distribution. Let $\mu_d(X_j; s_j, \theta)$ ($\theta = (w, a)$) be the median of the posterior distribution of μ_j. Given s_j and w, $\mu_d(X_j; s_j, w)$ is a weakly increasing function of X_j due to the thresholding property ($\mu_d(X_j; s_j, \theta)$ might be zero when x is close to zero). Thus, there is a unique threshold $t_j = t(s_j, \theta)$ for each observation j such that $\mu_d(X_j; s_j, \theta) = 0$ iff $X_j \leq t_j$.

Let $g(X_j; s_j, a)$ be the convolution of Laplace density $\gamma_a(\mu_j)$ and density of X_j and let $f_N(X_j; \mu_j, s_j)$ be the density of normal distribution with mean μ_j and standard deviation s_j. Then,

$$g(X_j; s_j, a) = \int \gamma_a(\mu_j) f_N(X_j; \mu_j, s_j) d\mu_j. \tag{1.5}$$

The marginal density of X_j is

$$X_j | s_j \sim (1-w) f_N(X_j; 0, s_j) + w g(X_j; s_j, a). \tag{1.6}$$

Weight w and Laplace parameter a are estimated through maximizing marginal log likelihood:

$$l(w, a | X_j, s_j, j = \{1, 2, \cdots, n\}) = \sum_{i=1}^{n} log((1-w) f_N(X_j; 0, s_j) + w g(X_j, s_j, a)) \tag{1.7}$$

subject to the constraint on weight w that $0 \leq t(s_j, w) \leq s_j \sqrt{2log(n)}$, $j = 1, 2, \cdots, n$. Weight estimate $\hat{w}$ helps to find the corresponding threshold $t_j = t(s_j, \hat{w})$. The constraint on weight is inherited from the case with noises of homogeneous variance in Johnstone and Silverman (2004)[1], where threshold range is from 0 to $\sqrt{2log(n)}$, the universal threshold for a sample of size n. This constraint is conservative in the sense of not over-shrinking the estimates. The authors mentioned that in their simulation with sparsest signals, the best results are obtained using universal threshold. However, it is far from clear why we need this constraint on thresholds and how this constraint improves the estimation. We will discuss the implication of this constraint in Section 3.

The next step would be to plug in $\hat{w}$ into the prior and estimate μ_j with either posterior median, posterior mean or other estimators.

2 Implementation

2.1 Derivation

In the following analysis, ϕ represents the density of a standard normal distribution, Φ the CDF of a normal distribution, and $\tilde{\Phi} = 1 - \Phi$.

The posterior distribution of μ under the above assumption is

$$\mu|x, s \sim (1 - w_{post})\delta_0(\mu) + w_{post} f_{post}(\mu|x, s, a) \tag{2.1}$$

in which w_{post} is the posterior probability of having a non-zero effect and f_{post} is the conditional density given the effect is non-zero.

The convolution $g(x; s, a)$ of a normal density with standard deviation s and a Laplace density with parameter a is

$$\begin{aligned} g(x; s, a) &= \int_{-\infty}^{\infty} f_N(x; \mu, s)\gamma_a(\mu)d_\mu \\ &= \frac{1}{2}ae^{\frac{a^2s^2}{2}}\left[e^{-ax}\Phi\left(\frac{x - s^2a}{s}\right) + e^{ax}\tilde{\Phi}\left(\frac{x + s^2a}{s}\right)\right]. \end{aligned} \tag{2.2}$$

The posterior weight can be expressed as

$$\begin{aligned} w_{post} &= P(\mu_j \neq 0|X_j, s_j) \\ &= \frac{w \cdot g(x; s, a)}{(1 - w)f_N(x; 0, s) + w \cdot g(x, s, a)} \\ &= \frac{w(\beta(x, s) + 1)}{1 + w\beta(x, s)} \end{aligned} \tag{2.3}$$

in which

$$\begin{aligned} \beta(x; s, a) &= \frac{g(x; s, a)}{f_N(x; 0, s)} - 1 \\ &= \frac{1}{2}as\left(\frac{\Phi(\frac{x-s^2a}{s})}{\phi(\frac{x-s^2a}{s})} + \frac{\tilde{\Phi}(\frac{x+s^2a}{s})}{\phi(\frac{x+s^2a}{s})}\right) - 1. \end{aligned} \tag{2.4}$$

The posterior distribution of μ conditional on a non-zero effect is therefore

$$f_{post}(\mu|x, s, a) = \frac{e^{-ax}\frac{1}{s}\phi(\frac{\mu-(x-s^2a)}{s})}{e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})}\mathbb{1}\{\mu > 0\} + \frac{e^{ax}\frac{1}{s}\phi(\frac{\mu-(x+s^2a)}{s})}{e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})}\mathbb{1}\{\mu < 0\} \tag{2.5}$$

- that is, a mixture of two truncated normal distributions that are symmetric with respect to the y axis. Define $TN(x;\mu,s,a,b)$ to be truncated normal distribution with location parameter μ, scale parameter s, minimum value a and maximum value b. Then,

$$f_{post}(\mu|x,s,a) = \lambda \cdot TN(\mu; x - s^2a, s, 0, +\infty) + (1-\lambda) \cdot TN(\mu; x + s^2a, s, -\infty, 0) \tag{2.6}$$

in which

$$\lambda = \frac{e^{-ax}\Phi(\frac{x-s^2a}{s})}{e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})} \tag{2.7}$$

is the probability of positive observations.

Suppose $x > 0$. The mean of the conditional distribution $f_{post}(\mu|x,s)$ is

$$\begin{aligned}\mu_m(x;s,a) &= \lambda \cdot (x - s^2a + \frac{\phi(\frac{x-s^2a}{s})}{\Phi(\frac{x-s^2a}{s})}s) + (1-\lambda) \cdot (x + s^2a - \frac{\phi(\frac{x+s^2a}{s})}{\tilde{\Phi}(\frac{x+s^2a}{s})}s) \\ &= \lambda(x - s^2a) + (1-\lambda)(x + s^2a).\end{aligned} \tag{2.8}$$

Thus, the posterior mean of $\mu|x,s$ is $w_{post} \cdot \mu_m(x;s,a)$.

The posterior median $\mu_d(x;s,\theta)$ will either have the same sign as the original value or be zero. Suppose $x > 0$ and, therefore, $\mu_d \geq 0$. Define

$$\begin{aligned}\tilde{F}_d(\mu|x,s) &= \int_{\mu}^{\infty} f_{post}(\mu|x,s)d\mu \\ &= \frac{e^{-ax}\tilde{\Phi}(\frac{\mu-(x-s^2a)}{s})}{e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})}\end{aligned} \tag{2.9}$$

for $\mu \geq 0$.

If there is a positive posterior median, μ_d has to satisfy $w_{post}\tilde{F}_d(\mu|x,s) = \frac{1}{2}$. Notice that there is no solution to this equation when $w_{post} < \frac{1}{2}$. The solution is non-positive when $w_{post} \geq \frac{1}{2}$ and $w_{post}\tilde{F}_d(0|x,s) \leq \frac{1}{2}$, which has opposite sign to what we expect, and the posterior median will be set to zero in this case. Thus, the posterior median $\mu_d(x;s,\theta)$ is the solution of

$$w_{post}\tilde{F}_d(\mu|x,s) = \frac{1}{2}, \text{ if } w_{post}\tilde{F}_d(0|x,s) > \frac{1}{2}. \tag{2.10}$$

Otherwise, the posterior median is zero.

With negative observations, we can follow the same analysis as above on their additive inverse and assign the sign of the corresponding observation to the posterior estimator. Therefore, we only discuss cases when observations are positive in the following analysis.

2.2 Algorithm

In this section, I revise and extend the algorithms in `R` package EbayesThresh to be compatible with data observed with noises of heterogeneous variance. The modifications of major `R` functions are listed as follows:

`beta.laplace` $\beta(x; s, a)$ in Equation (2.4) is calculated. In `R`, when $x > 35$, $\phi(x)$ tends to be zero, and $\Phi(x)$ tends to be 1 due to a limit of numerical accuracy `R` can reach. $\frac{\Phi(x)}{\phi(x)}$, thus, goes to infinity, which will cause a problem when maximizing the marginal log likelihood function. Meanwhile, $\frac{\tilde{\Phi}(x)}{\phi(x)}$ produces a missing value issue. In practice, $\frac{\tilde{\Phi}(x)}{\phi(x)}$ is approximately equal to $\frac{1}{x}$ when $x > 35$ (See Silverman and Johnstone (2005)[2]). In the original code of EbayesThresh, the missing value issue is solved by the above approximation. To avoid issues due to infinite values, I approximate $\frac{\Phi(x)}{\phi(x)}$ by $\frac{\Phi(35)}{\phi(35)}$ when $x > 35$.

`postmean.laplace` The posterior mean $w_{post} \cdot \mu_m(x; s, a)$ is calculated. $\mu_m(x; s, a)$ in Equation (2.8) can be written as

$$\mu_m(x; s, a) = x - as^2 \left(\frac{2\Phi(\frac{x-s^2a}{s})}{\Phi(\frac{x-s^2a}{s}) + e^{2ax}\tilde{\Phi}(\frac{x+s^2a}{s})} - 1 \right) \tag{2.11}$$

for calculation in `R`.

When s or a is large, $\Phi(\frac{x-s^2a}{s})$ and $\tilde{\Phi}(\frac{x+s^2a}{s})$ might go to zero, which will generate missing value of the posterior mean. To avoid issues due to missing values, I approximate $\Phi(x)$ by $\Phi(-35)$ when $x < -35$ and $\tilde{\Phi}(x)$ by $\tilde{\Phi}(35)$ when $x > 35$.

`postmed.laplace` Posterior median in Equation (2.10) is calculated. We can derive the posterior median from Equation (2.10) if posterior median is positive:

$$\begin{aligned}
& w_{post}\tilde{F}_d(\mu|x, s) = \frac{1}{2} \\
\Leftrightarrow & \frac{e^{-ax}\tilde{\Phi}(\frac{\mu-(x-s^2a)}{s})}{e^{-ax}\Phi(\frac{x-s^2a}{s}) + e^{ax}\tilde{\Phi}(\frac{x+s^2a}{s})} = \frac{(1-w)\frac{1}{s}\phi(\frac{x}{s}) + w \cdot g(x; s, a)}{2wg(x; s, a)} \\
\Leftrightarrow & \tilde{\Phi}(\frac{\mu-(x-s^2a)}{s}) = (aw)^{-1}(1 + w\beta(x; s, a))\frac{1}{s}\phi(\frac{x-s^2a}{s}) \\
\Leftrightarrow & \mu = x - s^2a - s\Phi^{-1}(z(x, s))
\end{aligned} \tag{2.12}$$

in which

$$z(x, s) = a^{-1}(w^{-1} + \beta(x; s, a))\frac{1}{s}\phi(\frac{x-s^2a}{s}). \tag{2.13}$$

As $\frac{x-s^2a}{s} \to \infty$, $z(x, s)$ converges to $\frac{1}{2}$. This approximate value of $z(x, s)$ will be used when $\frac{x-s^2a}{s}$ is larger than 25 lest $\beta(x, s)$ become infinity and $\phi(\frac{x-s^2a}{s})$ go to zero. Notice that

when $w_{post} < \frac{1}{2}$, $z(x,s)$ can be larger than 1. Therefore,

$$\mu_d(x,s) = \max\{0, x - s^2a - s\Phi^{-1}(\min\{1, z(x,s)\}). \tag{2.14}$$

`tfromw(prior='laplace')` When $x - s^2a - s\Phi^{-1}(z(x,s)) < 0$, the posterior median will be set to zero. Thus, there is a posterior median threshold $t(s,w,a)$ such that the estimate of μ is zero whenever $|x| < t(s,w,a)$. This threshold, t, satisfies

$$0 = t - s^2a - s\Phi^{-1}(z(t,s)) \tag{2.15}$$

$$\Leftrightarrow \Phi(\frac{t - s^2a}{s}) = z(t,s). \tag{2.16}$$

Given s, w and a, the left hand side of Equation (2.16) goes to 1 and the right hand side goes to $\frac{1}{2}$ as we have discussed above when $t \to \infty$. Besides, $z(0,s) \geq \Phi(\frac{-s^2a}{s})$ and the equality holds iff $w = 1$. Thus, either there will be at least one solution approximately in the region $(0, 25s + s^2a)$ or the threshold should be zero. The solution of Equation (2.16) can be found by binary search in the interval $[0, 25s + s^2a]$. In contrast to the original analysis with homogeneous variance in Johnstone and Silverman (2004)[1], the threshold might vary by observation j since the standard deviation of observations might be different, which means t_i is not necessarily equal to t_j if $s_i \neq s_j$.

`wfromt(prior='laplace')` The weight w can be derived in terms of t, s and a from the above formula:

$$w(t,s,a) = \left(as\frac{\Phi(\frac{t-s^2a}{s})}{\phi(\frac{t-s^2a}{s})} - \beta(t;s,a) \right)^{-1}. \tag{2.17}$$

w is well defined in $[0,1]$ and monotonically declines with t given s and a. Note that w defined above might be different for different threshold and standard deviation.

`wandafromx` w and a will be estimated by maximizing marginal log likelihood in the empirical bayes sense. The marginal log likelihood is as mentioned in Equation (1.7)

$$l(w,a|X_j, s_j, j = \{1,2,\cdots,n\}) = \sum_{i=1}^{n} log((1-w)f_N(X_j;0,s_j) + wg(X_j;s_j,a)) \tag{2.18}$$

which is the same as to maximize

$$\sum_{i=1}^{n} log(1 + w\beta(x_i;s_i,a)). \tag{2.19}$$

We can either search the optimal w and a directly by maximizing the marginal log likelihood above or search with constraints on threshold. If searching with constraints, each

threshold t_j is upper bounded by universal threshold, $0 < t_j < s_j\sqrt{2log(n)}$, which does not allow a very large probability mass at zero. Thus, w is constrained by the intersection of the range of $w(t_j, s_j, a)$ in Equation (2.17) in terms of domain of t_j, s_j and a, for $j = 1, 2, \cdots, n$:

$$w \in W(a, s_j, j = \{1, 2, \cdots, n\}) = \bigcap_{j=1}^{n} \{w(t_j, s_j, a) : 0 < t_j < s_j\sqrt{2log(n)}\}. \tag{2.20}$$

Thus, the optimization problem is

$$\begin{aligned} &\max_{w,a} \sum_{i=1}^{n} log(1 + w\beta(x_i; s_i, a)) \\ s.t.\ w \in W(a, s_j, j = \{1, 2, \cdots, n\}) &= \bigcap_{j=1}^{n} \{w(t_j, s_j, a) : 0 < t_j < s_j\sqrt{2log(n)}\}. \end{aligned} \tag{2.21}$$

$W(a, s_j, j = \{1, 2, \cdots, n\})$ can be simplified in terms of the monotonically negative relationship between w and t given s and a. Since t_j lies in the interval $[0, s_j\sqrt{2log(n)}]$,

$$W(a, s_j, j = \{1, 2, \cdots, n\}) = [\max_{j=1,2,\cdots,n} w(s_j\sqrt{2log(n)}, s_j, a), 1]. \tag{2.22}$$

`wfromx` If value of a is provided, a general method to find the optimal w would be to find the root of $S(w) = 0$ (Silverman and Johnstone (2005)[2]) with

$$S(w) = l'(w) = \sum_{i=1}^{n} \frac{\beta(x_i; s_i, a)}{1 + w\beta(x_i; s_i, a)} \tag{2.23}$$

in the searching region $[\max_{j=1,2,\cdots,n} w(s_j\sqrt{2log(n)}, s_j, a), 1]$ as discussed above. $S(w)$ is a decreasing function of w so the root can be found through a binary search. If there is no root within the region, a boundary solution will be taken[2].

3 Experiment

3.1 Accuracy of Parameter Estimation

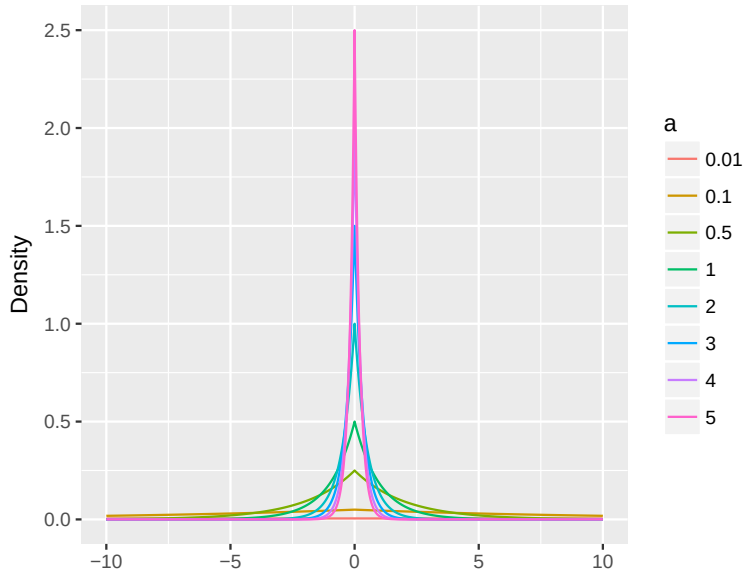
We first test the efficiency of EbayesThresh by estimating Laplace parameter a and non-null weight w under the actual model. Consider a data sequence of which each effect is drawn from

$$\mu \sim (1-w)\delta_0(\mu) + w\gamma_a(\mu) \tag{3.1}$$

as in Equation (1.3). The true effect μ_i of subject i is observed with noise from a normal distribution $N(0, s_i^2)$, with $s_i^2 \sim \chi_1^2$. The goal is to see if estimates of w and a will be close to the true values.

The authors mentioned a reasonable range for a would be between 0.04 and 3[2]. From Figure 1, we can find that when a is small the variation is too large to provide effective estimation; when a is large, the variation is very small so that the effect perhaps should be treated as null. We will consider a to be between 0.04 and 3 here as well.

Figure 1: Laplace density with mean zero and different scale parameters



We estimate w and a from a data sequence of size 10,000 sampled through the above process. Estimation results $\hat{w}$ and $\hat{a}$ of different true values of w and a are shown in Figure 2. Panel (a) - (b) show estimation with constraint on weight as in Equation (2.22) and Panel (c) - (d) show estimation without constraint, in which case $w \in [0, 1]$. The result shows that estimates of w and a tends to be larger than the true values. The two results with/without constraints are very close to each other, ceteris paribus, when weight is large. However, both

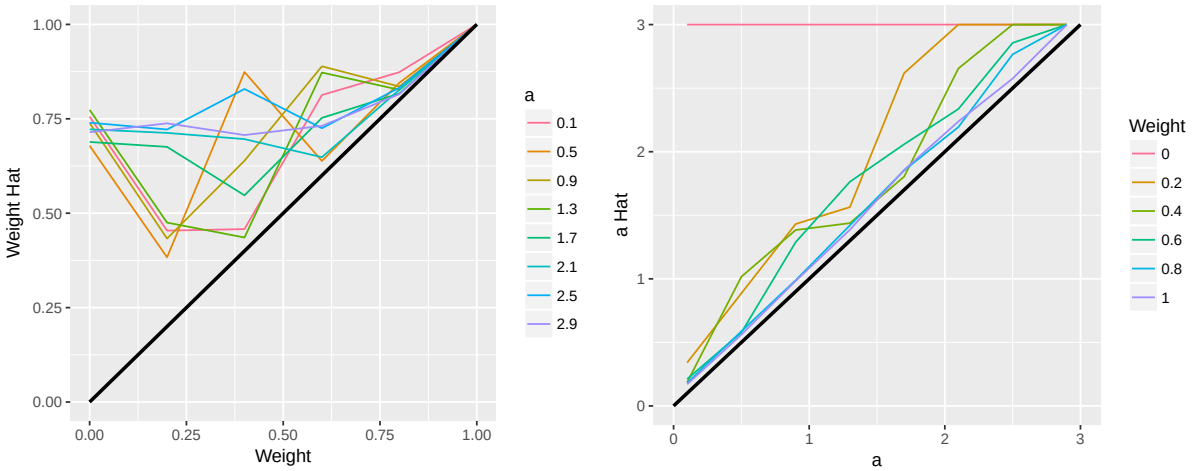
the weight and Laplace parameter are overestimated when the true value of weight is small due to the constraint.

Figure 2: Estimation of weight w and Laplace parameter a in the actual model

Estimates of weight are plotted over true weight given different Laplace parameters a and estimates of a are plotted over true parameters given different weight. Panel (a) - (b) shows estimation with constraint on weight while Panel (c) - (d) shows estimation without constraint on weight. The black line is $y = x$.

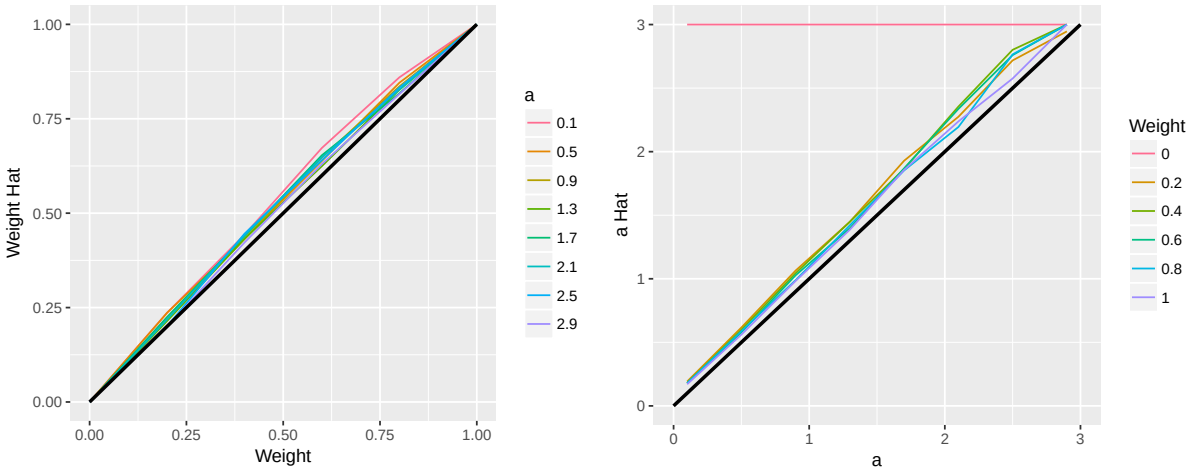
(a) Weight estimates w/ constraint

(b) Laplace parameter estimates w/ constraint



(c) Weight estimates w/o constraint

(d) Laplace parameter estimates w/o constraint



3.2 Performance

Next, we discuss the performance of EbayesThresh with heterogeneous variance and constraint on weight when dealing with data with noises from heterogeneous variance. Consider a data sequence of 2,000 observations with m null effects and $2,000 - m$ effects drawn from $N(0, 1)$. Each true effect μ_i is observed with noise drawn from a normal distribution $N(0, s_i^2)$, with $s_i^2 \sim \chi_1^2$. In this case, the expected mean squared error is 1 if using naive estimator -

the observations, X_i -

$$\begin{aligned} E(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_i)^2) &= E(\frac{1}{n}\sum_{i=1}^{n}E((X_i-\mu_i)^2|s_i)) \\ &= E(\frac{1}{n}\sum_{i=1}^{n}s_i^2) \\ &= 1. \end{aligned} \tag{3.2}$$

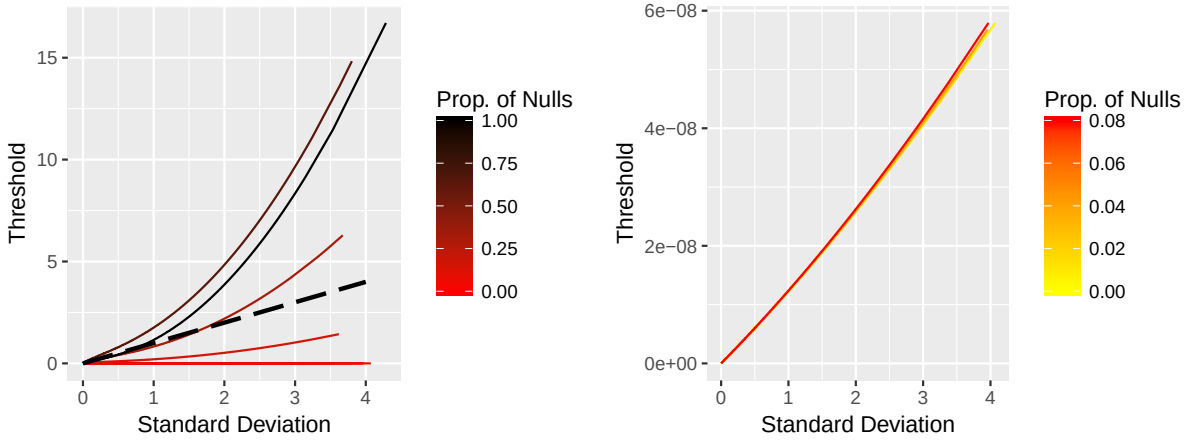
Here we consider the number of null effects $m = 0, 5, 10, 20, 40, 80, 160, 320, 640, 1280,$ and 2000 in the following analysis.

The performance of posterior median μ_d and posterior mean μ_m as estimators of true effects with noises of heterogeneous variance is shown in Figure 3. Observations with larger standard deviation tend to have posterior estimates closer to zero. It is intuitively reasonable since larger standard deviation means more uncertainty and less information and, thus, posterior estimates will be closer to the prior, which is symmetric by y axis. It can also be observed through the increasing gap between threshold and standard deviation with standard deviation shown in Figure 4. When the proportion of null effects is large, the threshold tends to be much larger than the standard deviation and, thus, observations with larger standard deviation given large proportion of null effects tend to have posterior median of zero.

Figure 4: Threshold versus standard deviation

Estimated threshold is plotted versus standard deviation for different proportion of null effects. Panel (b) zooms in those curves of small proportion of null effects. The dashed line in panel (a) shows $y = x$.

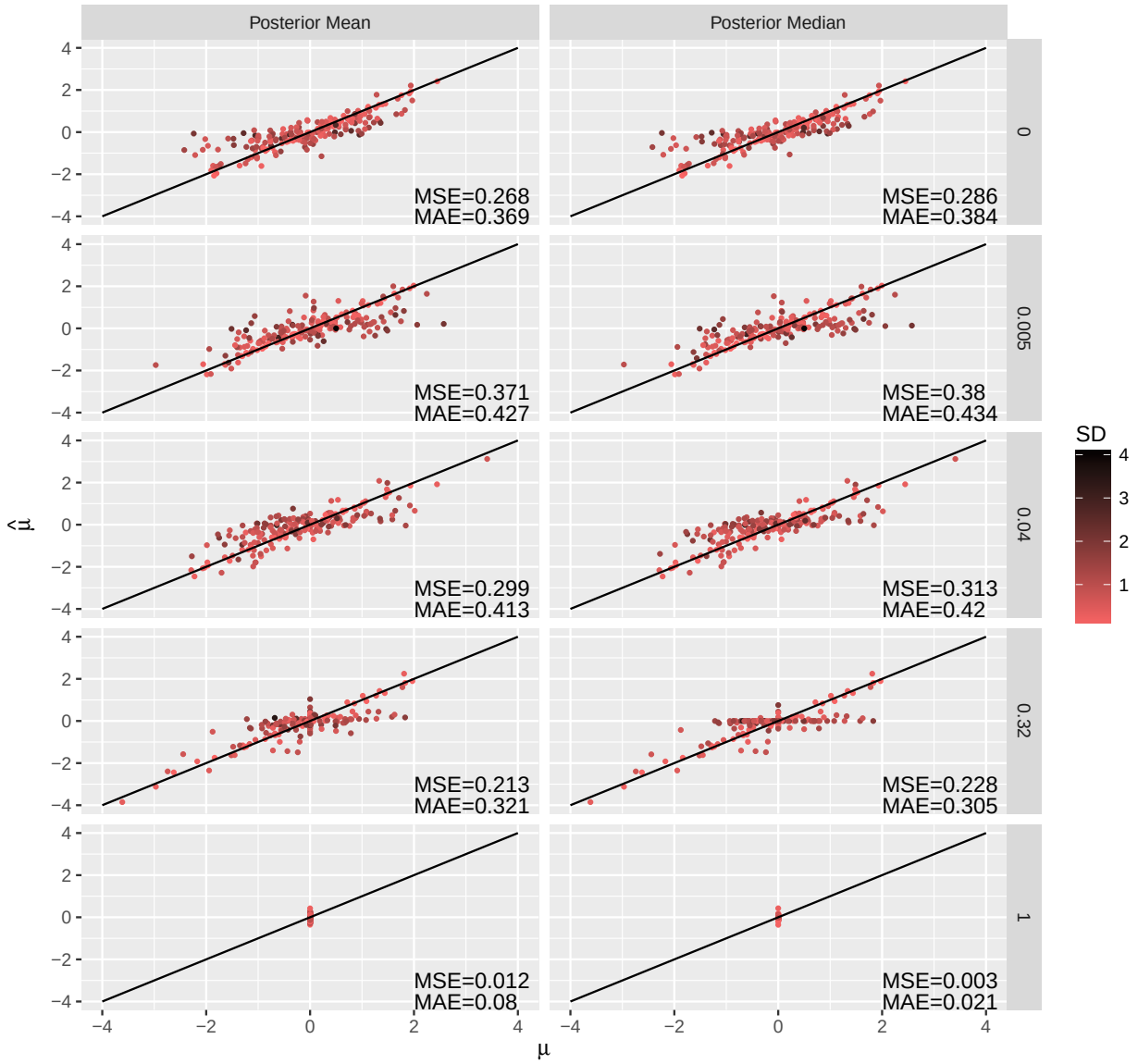
(a) For all groups of proportion of nulls (b) For groups with small proportion of nulls



To see how easing the restriction of homogeneous standard deviation improves the estimation, we compare mean squared error (MSE) and mean absolute error (MAE) over different proportion of null effects with different restrictions on standard deviation. Three different standard deviations are passed to the model: i) homogeneous standard deviation measured

Figure 3: Posterior estimates of true effects

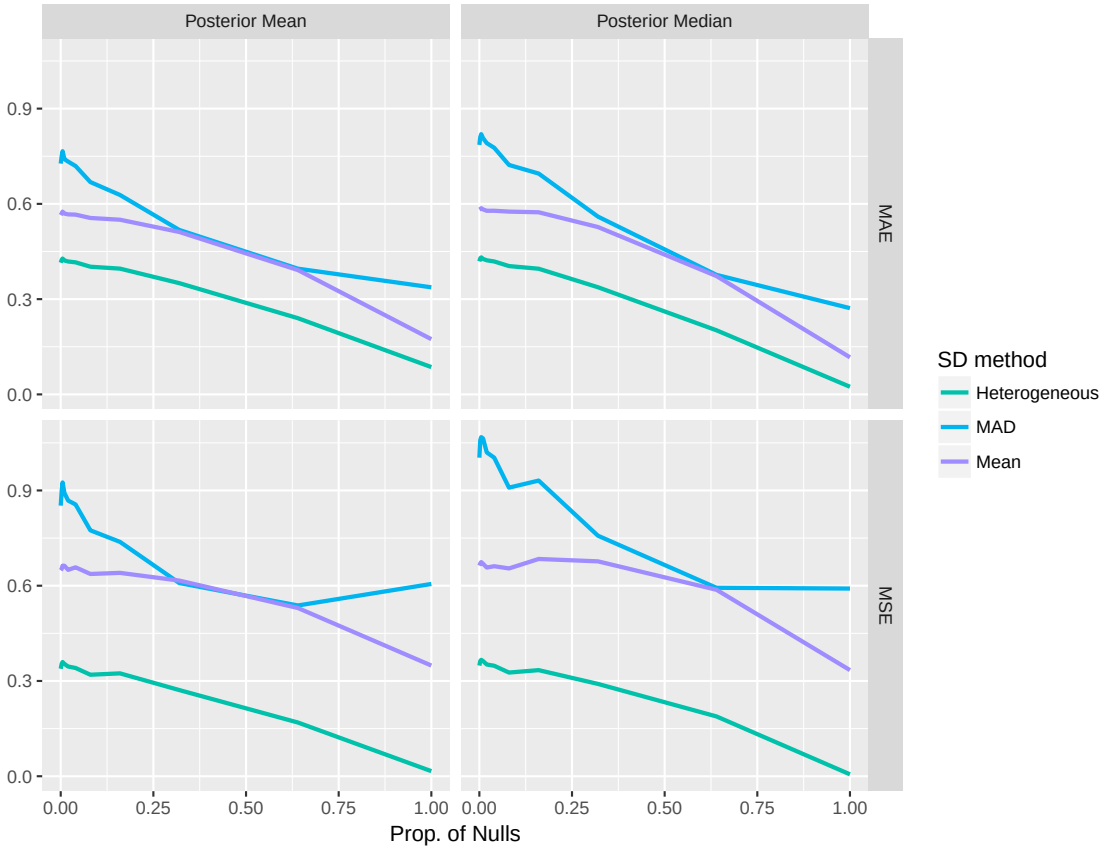
The posterior mean and median are compared with true effects across different proportion of null effects. Non-zero effects are initially drawn from $N(0,1)$. True effects are observed with noise from $N(0,s_i^2)$, $s_i^2 \sim \chi_1^2$. The black line shows $y=x$. The estimation results of $m=0,10,80$, and 640 (corresponding to proportion $0, 0.005, 0.04$ and 0.32) are plotted every 10 data points to make the trend clearer. For example, the sequence of the $10^{th}, 20^{th} \cdots$ of the observations are chosen. Mean squared error(MSE) and mean absolute error(MAE) are shown in each panel.



by median absolute deviation (MAD) of the observations, ii) homogeneous standard deviation measured by mean of the true standard deviations, and iii) the true heterogeneous standard deviations. For each proportion of null effects and each standard deviation, we calculate MSE and MAE 100 times and then take the average. Figure 5 shows results for both posterior mean and posterior median as posterior estimator.

Figure 5: Comparison of models with/without restriction on homogeneous standard deviation in terms of MSE/MAE

MAE and MSE of posterior mean and median under different model assumptions are plotted over proportion of null effects. For example, the top-left panel plots the MAEs using posterior mean as posterior estimator of three models with/without restriction on homogeneous standard deviation.



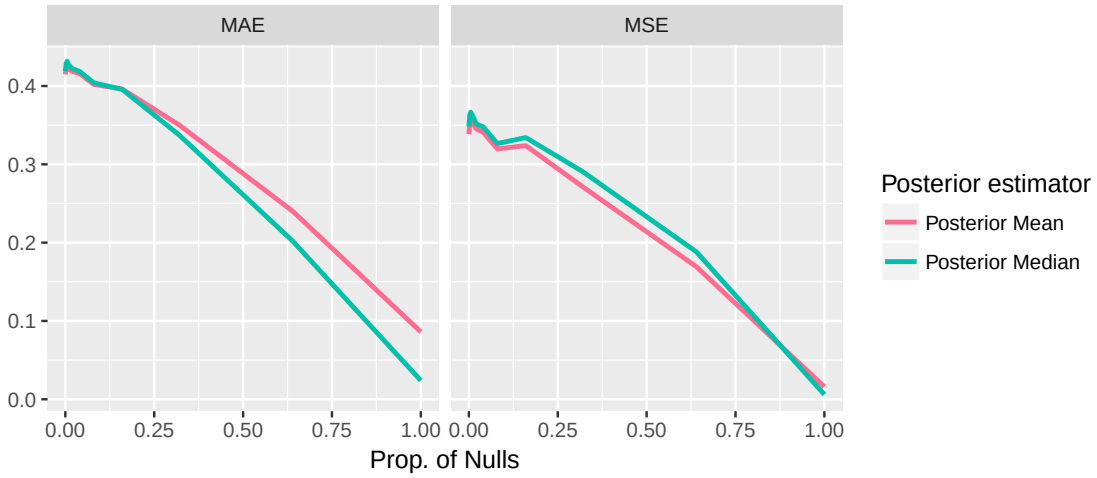
The expected value of mean squared error of naive estimator is 1 (Equation (3.2)) and we can see all methods provide a MSE smaller than 1 using posterior mean as estimator at the bottom-left panel in Figure 5. The model with heterogeneous standard deviation excels at controlling error measured in terms of MSE. The MSE of estimates under heterogeneous standard deviation are much smaller than those with restriction of a homogeneous standard deviation assumption. Using mean of the true standard deviations as the standard deviation under homogeneity assumption provides more information than none so the corresponding MSE is lower than that using MAD of the observations as the standard deviation. We also calculate MSE based on posterior median and MAE based on posterior mean and median,

and the corresponding plots are shown in Figure 5. We can find similarly posterior median provides much smaller MAE under the model with heterogeneous standard deviation.

Another thing of interest is whether posterior mean would be more robust than posterior median in terms of MSE and posterior median performs better than posterior mean in terms of MAE. Figure 6 compares the MAE and MSE of posterior mean and median under model with heterogeneous standard deviation. Posterior median performs much better than posterior mean in terms of MAE when the proportion of null effects is large. Meanwhile, posterior mean outperforms posterior median using error measure of MSE when the proportion of null effects is small.

Figure 6: Comparison of performance of posterior mean/median under models with heterogeneous variance in terms of MSE(MAE)

MAE and MSE of posterior mean and median under model with heterogeneous standard deviation are plotted over proportion of null effects. For example, the left panel plots the MAEs of the two posterior estimators under models without restriction on homogeneous standard deviation.



We now consider a more extreme case to see if the method is still effective when there are more observations with poor measurement precision (standard deviation of noises). We simulate 200 data points with true effects from

$$\mu \sim 0.5\delta_0(\mu) + 0.5N(\mu; 0, 1). \tag{3.3}$$

The first half of the observations have precise measurement (standard error $s_i = 1$) and the second half are observed with large noises (standard error $s_i = 10$).

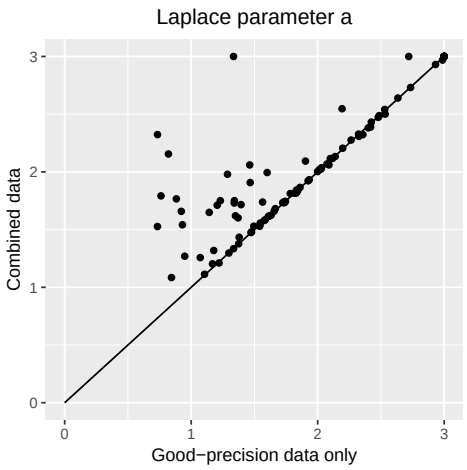
In this example, little information is provided by data points with poor precision and we would expect an effective method to provide a same result as if there are no poor observations. Repeat the above simulation for 100 times. Figure 7 compares estimation of Laplace parameter a, weight w and MSE of combined data with both good and poor precision versus data with good precision only. The estimates of Laplace parameter a tend to be larger when

we have both precise and poor data points. However, the estimations of weight and MSE are very similar in both cases, which shows the effectiveness of the method.

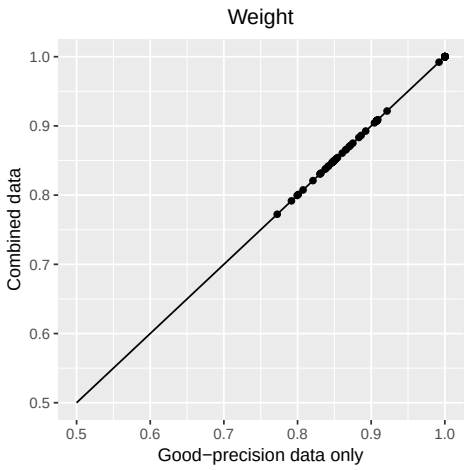
Figure 7: Robustness of varying precision

Panel (a) - (c) shows the estimates of Laplace parameter a, weight w and MSE from combined data and good-precision data only.

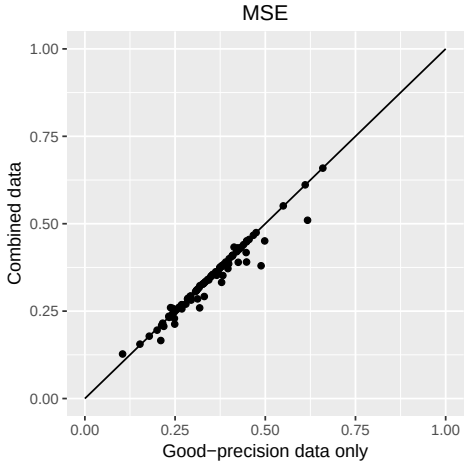
(a)



(b)



(c)



To better understand the overestimation of Laplace parameter a in Figure 7, we repeat the above analysis without the constraint on weight. In this case, the estimates of Laplace parameter, weight and MSE are nearly the same using combined data or good-precision data only (Figure 8). One thing worth mentioning is the potential positive relation between the estimates of weight and Laplace parameter. For example, suppose the weight is overestimated. In order to compensate for the bias due to underestimation of the quantity of null effects, the estimated variance of Laplace distribution is expected to be smaller so that it is more likely to observe effects close to zero from the Laplace distribution. Therefore, a is expected to be larger as the variance of Laplace distribution defined in Equation (1.4) is $\frac{2}{a^2}$.

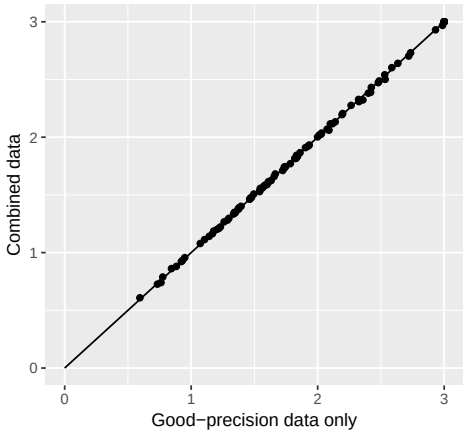
The scattering estimated weights in $[0, 1]$ might be a result of multiple optimal solutions of a combination of a and w when maximizing the marginal likelihood, due to the potential positive relation between a and w.

When adding lower bound on the weight, the weight estimates might take the lower boundary value when the corresponding estimates without constraint are small. However, the different shapes of the marginal log likelihood functions of combined data and good-precision data given a same weight might cause the Laplace parameter estimates to be larger for combined data. This in some respects explains the similarity of weight estimates and the difference of Laplace parameter estimates in Figure 7 when estimating with constraint on weight.

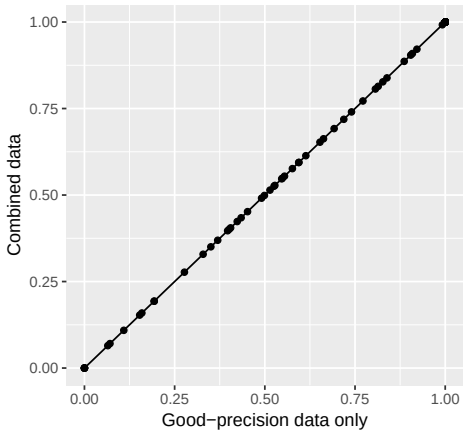
Figure 8: Robustness of varying precision w/o constraint on weight

Panel (a) - (c) shows the estimates of Laplace parameter a, weight w and MSE from combined data and good-precision data only.

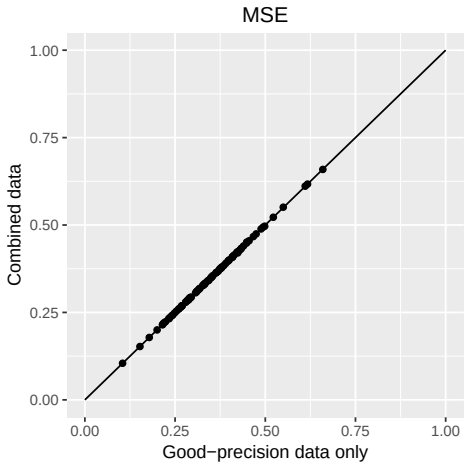
(a)



(b)



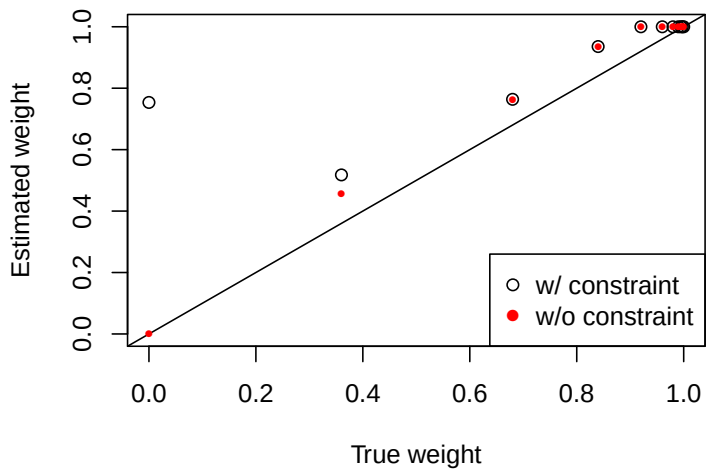
(c)



3.3 Discussion of Constraint on Weight

One question would be how the constraint on weight (Equation (2.22)) influences the weight estimation, and indirectly the MSE. Due to the constraints on threshold, we would expect the estimated weight not to be smaller than the true weight. We have already shown the potential influence on weight and Laplace parameter estimation in Figure 2. Since the constraint does not allow a too small weight, it reduces accuracy by overestimating weight of non-zero effects when the proportion of null effects is large. Figure 9 shows that the weight estimation is hardly biased when the non-zero weight is large but deviates from the true weight otherwise.

Figure 9: Comparison of estimated weight with/without constraints on weight



However, when the proportion of null effects is large, the MSE is rather small and the increase in error due to a biased weight estimator is negligible (see the bottom-left panel in Figure 10). The same result is observed for MAE with posterior median at the top-right panel in Figure 10. We have shown in Figure 4 that the threshold tends to be larger than standard deviation when the proportion of null effects is large. Thus, even if the weight is poorly estimated, the shrinkage might still works efficiently. The MSE calculated based on posterior median and the MAE based on posterior mean are also shown in Figure 10.

Figure 11 shows the proportion of MSE(MAE) with constraint on weight to that without constraint. We can find MSE(MAE) is hundreds of times larger with constraint on weight when the proportion of nulls is large, though the absolute value of MSE(MAE) doesn't differ much with or without constraint.

Figure 10: Comparison of MSE(MAE) with/without constraint on weight

MSE(MAE) of posterior mean and median are calculated under model with/without constraint on weight. For example, the top-left panel shows the MAE using posterior mean as posterior estimator under model with/without constraint on weight.

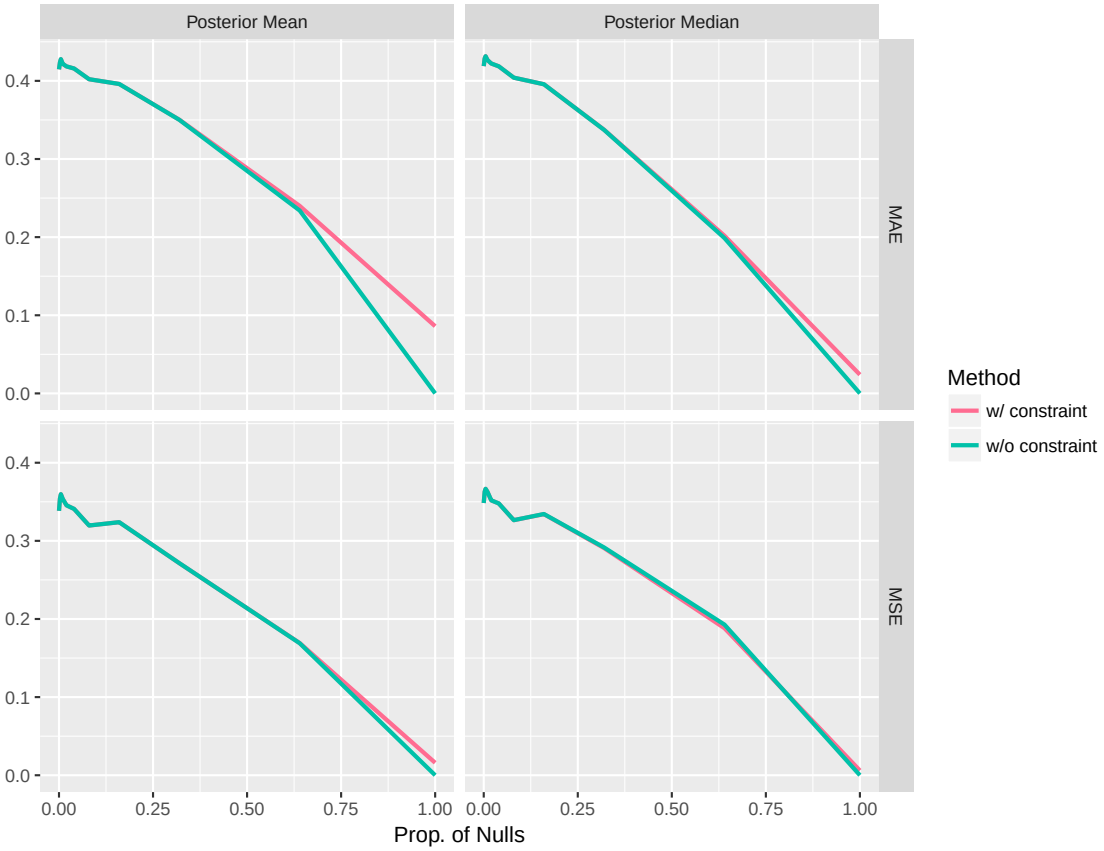
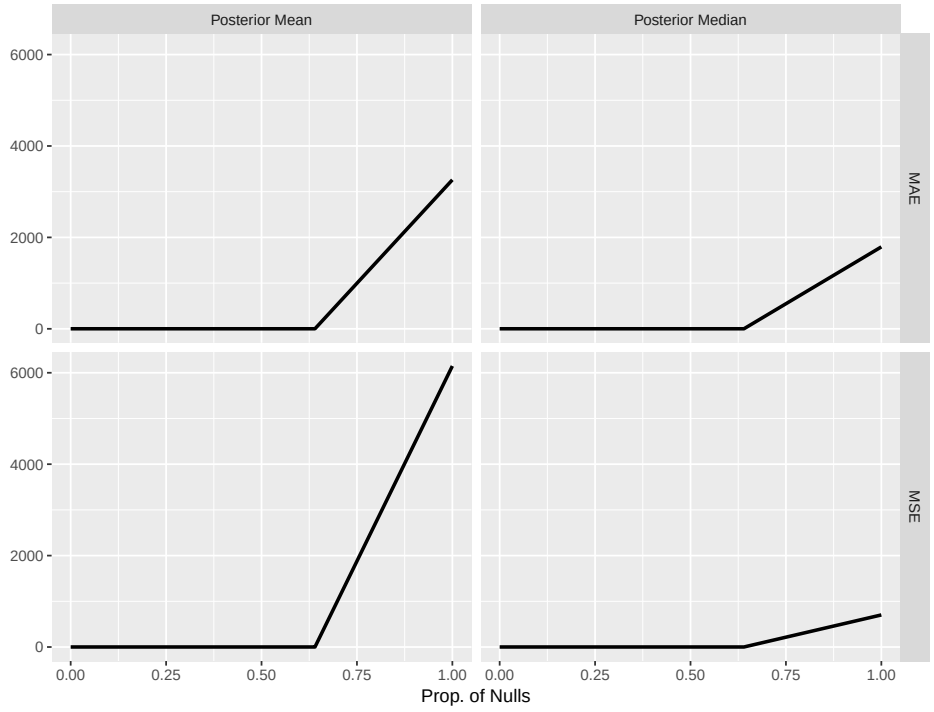


Figure 11: Proportion of MSE(MAE) with constraint on weight to that without constraint



4 Conclusions

We expand the model under assumption of noises of homogeneous variance in Johnstone and Silverman (2004)[2] into a heterogeneous case. The model with heterogeneous standard deviation gives estimates close to the true values when the non-null weight is large under the actual model. This ease of restriction on standard deviation improves the model estimation in terms of MSE and MAE. The MSE tends to be smaller for data with larger proportion of null effects. Posterior mean is more robust than posterior median in terms of MSE and posterior median outperforms posterior mean in terms of MAE.

We also discuss the potential bias due to the constraint on weight in the setting of EbayesThresh. The range of the threshold is assumed to be $[0, \sqrt{2log(n)}]$ in the original homogeneous case and is extended to $[0, s_i\sqrt{2log(n)}]$ for each threshold t_i given heterogeneous standard deviation. This constraint is shown to largely overestimate weight of non-zero effects when the proportion of null effects is large because the upper bound on threshold does not allow a too small weight. However, this constraint does not influence MSE and MAE much in terms of absolute value.

References

[1] Johnstone, Iain M., and Bernard W. Silverman (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics* (2004): 1594-1649.

[2] Silverman, Bernard W., and Iain Johnstone (2005). EbayesThresh: R Programs for Empirical Bayes Thresholding. *Journal of Statistical Software* 12.08 (2005).

[3] Stephens, Matthew (2016). False discovery rates: a new deal. *Biostatistics (2016) : kxw041.*