Summary and discussion of: "Flexible Signal Denoising via Flexible Empirical Bayes Shrinkage"

RUIZHAO HUANG

1   Summary

1.1   Empirical Bayes for Normal Mean

This section corresponds to the concepts covered in the lecture. Let us assume we have a set of observations, denoted as $\boldsymbol{z} = (z_1, \ldots, z_N)^T$, where:

$$z_i | \mu_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_i, 1), \quad i = 1, \ldots, N$$

In a compact matrix form, we have:

$$\boldsymbol{z} | \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_N)$$

Here, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)^T$ are unknown, and want to estimate. If we assume homoskedastic Gaussian prior to $\boldsymbol{\mu}$, such that:

$$\mu_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, N$$

In a compact matrix form, we have:

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_N)$$

Under these assumptions, the marginal distribution of $\boldsymbol{z}$ will be:

$$\boldsymbol{z} \sim \mathcal{N}\big(\boldsymbol{\mu}, (1 + \sigma^2)\mathbf{I}_N\big)$$

The general Empirical Bayes method to estimate $\boldsymbol{\mu}$ in this case can be summarized in the following two steps:

a) Get the estimate of σ^2 by maximize the marginal likelihood of $\boldsymbol{z}$, which is adaptively estimated from data.

b) Incorporate the estimated value of σ^2 and derive the posterior distribution of $\boldsymbol{\mu}$. The estimation of $\boldsymbol{\mu}$ can be obtained using either the posterior mean or median.

It is worth to note that in this case, we can leverage the distributional information of $\boldsymbol{z}$ and its relationship with $(1 + \sigma^2)$, such that $\sum_{i=1}^{N} \frac{z_i^2}{1+\sigma^2} \sim \chi_N^2$. Consequently, the order of the above procedure is actually reversed. However, for consistency with subsequent cases, I have presented it in this order.

1.2 Empirical Bayes for Sparse Normal Mean

Next, within the framework established in the previous section, we consider the scenario where $\boldsymbol{\mu}$ is assumed to be sparse. To tackle this case, I will outline the method proposed by the author in another paper called `ASH` [1] — this method serves as a fundamental component of the approach introduced in this paper, which is called `SMASH` [2].

The major differences of `ASH` with previous section are as following:

a) Instead of assuming a homoskedastic Gaussian prior, we consider the individual components of $\boldsymbol{\mu}$ to have a unimodal (mixture) prior, such that:

$$\mu_i \overset{\text{indep}}{\sim} \pi_0 \delta_0(\mu_i) + \sum_{i=1}^{K} \pi_k \mathcal{N}(\mu_i; 0, \sigma_k^2), \quad i = 1, \ldots, N, \quad \sum_{k=0}^{K} \pi_k = 1 \tag{1}$$

where δ_0 represents the point mass at 0. This prior choice intuitively captures our knowledge regarding the sparsity of $\boldsymbol{\mu}$, allowing for the possibility that some μ_i values are exactly 0 with a certain probability π_0.

b) Instead of assuming $z_i | \mu_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_i, 1)$, we consider the following assumption:

$$z_i | \mu_i \overset{\text{indep}}{\sim} \mathcal{N}(z_i; \mu_i, \hat{s}_i^2) \implies \boldsymbol{z} | \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \hat{\boldsymbol{S}}), \; \hat{\boldsymbol{S}} = \text{diag}(\hat{s}_1^2, \ldots, \hat{s}_N^2) \tag{2}$$

where $\hat{s}_i^2$ represents the estimated variance of $z_i | \mu_i$, and its value is provided.

c) The parameters $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)$ are estimated using a grid search approach during the maximization of the marginal log-likelihood of $\boldsymbol{z}$. The grid resolution is set to be:

$$\sigma_{min} = \frac{\min\{\hat{s}_i\}}{10}, \quad \sigma_{max} = 2\sqrt{\max\left\{\hat{z}_i^2 - \hat{s}_i^2\right\}}, \quad \sigma_k = \sqrt{2}\sigma_{k-1}$$

Under these assumptions, the marginal density of z_i are mixture Normal, and can be expressed as:

$$p(z_i) = \sum_{k=0}^{K} \pi_k \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_k^2\right), \quad \sigma_0^2 = 0, \quad i = 1, \ldots, N$$

The marginal log-likelihood of $\boldsymbol{z}$ can be defined as:

$$l(\boldsymbol{z}) = \sum_{i=1}^{N} \ln\left(\sum_{k=0}^{K} \pi_k \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_k^2\right)\right)$$

It is worth noting that since we employ a grid search to estimate σ_k^2, which treats σ_k^2 as fixed, the function $L(\boldsymbol{z}, \boldsymbol{\pi})$ is concave with respect to $\boldsymbol{\pi}$. This is a result of the concavity of the logarithm function and the fact that a non-negative weighted concave function remains concave.

Thus, the optimal solution for $\boldsymbol{\pi}$ can be obtained by solving the following convex optimization problem:

$$\begin{array}{ll} \underset{\boldsymbol{\pi}}{\text{Maximize}} & \sum_{i=1}^{N} \ln\left(\sum_{k=0}^{K} \pi_k \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_k^2\right)\right) \\ \text{subject to} & \mathbf{1}^T \boldsymbol{\pi} = 1, \quad \boldsymbol{\pi} \succeq \mathbf{0} \end{array}$$

However, this paper considers the following penalized convex optimization problem:

$$\text{Maximize}_{\boldsymbol{\pi}} \quad \sum_{i=1}^{N} \ln\left(\sum_{k=0}^{K} \pi_k \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_k^2\right)\right) + \sum_{k=0}^{K} (\lambda_k - 1) \ln \pi_k$$
$$\text{subject to} \quad \mathbf{1}^T \boldsymbol{\pi} = 1, \quad \boldsymbol{\pi} \succeq \mathbf{0}$$

The default settings are $\lambda_0 = 10$, and $\lambda_k = 1$ for $k > 0$. It is important to note that under this setting, $\lambda_k - 1 = 0$ for $k > 0$, and since we are maximizing the objective function, the optimization process will tend to push π_0 away from 0, which can be interpreted as an over-estimation of π_0 to encourage sparsity.

Alternatively, we can use the EM algorithm to estimate $\boldsymbol{\pi}$, as the marginal density of z_i follow the mixture Normal distribution. To do this, let's introduce a latent variable $\boldsymbol{x}_i = (x_{i0}, \ldots, x_{iK})$ defined as:

$$\mathbb{P}(x_{ik} = 1) = \pi_k, \quad 0 \leqslant \pi_k \leqslant 1, \quad \sum_{k=0}^{K} \pi_k = 1$$

Consequently, we can express the joint distribution of $\boldsymbol{x}$ and z_i compactly as follows:

$$p(\boldsymbol{x}_i, z_i) = \prod_{k=0}^{K} \left(\pi_k \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_k^2\right)\right)^{x_{ik}}, \quad i = 1, \ldots, N$$

Next, the complete log-likelihood of $(\boldsymbol{x}, \boldsymbol{z})$ is given by:

$$\ln p(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{N} \sum_{k=0}^{K} x_{ik} \left(\ln \pi_k + \ln \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_k^2\right)\right)$$

Suppose we have the previous parameter $\boldsymbol{\pi}^{\text{old}}$, then the expectation of the complete log-likelihood with respect to the posterior distribution of the latent variable $\boldsymbol{x}$ at $\boldsymbol{\pi}^{\text{old}}$ is:

$$\mathbb{E}_{\boldsymbol{x}|\mathbf{z}, \boldsymbol{\pi}^{\text{old}}}\left(\ln p(\boldsymbol{x}, \boldsymbol{z})\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{x}|\mathbf{z}, \boldsymbol{\pi}^{\text{old}}}(x_{ik}) \left(\ln \pi_k + \ln \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_k^2\right)\right)$$

Since x_{ik} is a Bernoulli random variable, we have:

$$\gamma(x_{ik}) := \mathbb{E}_{\boldsymbol{x}|\mathbf{z}, \boldsymbol{\pi}^{\text{old}}}(x_{ik}) = \mathbb{P}_{\boldsymbol{x}}\left(x_{ik} = 1 | z_i, \boldsymbol{\pi}^{\text{old}}\right)$$
$$= \frac{\pi_k \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_k^2\right)}{\sum_{j=0}^{K} \pi_j \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_j^2\right)}$$

Furthermore, in the maximization step, we need to incorporate the penalty term, resulting in the following convex optimization problem:

$$\text{Maximize}_{\boldsymbol{\pi}} \quad \sum_{i=1}^{N} \sum_{k=0}^{K} \gamma(x_{ik}) \left(\ln \pi_k + \ln \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_k^2\right)\right) + \sum_{k=0}^{K} (\lambda_k - 1) \ln \pi_k$$
$$\text{subject to} \quad \mathbf{1}^T \boldsymbol{\pi} = 1, \quad \boldsymbol{\pi} \succeq \mathbf{0}$$

To solve this convex optimization problem, note that we can optimize each coordinate individually (for each k) since they are separable. We can utilize Lagrangian multipliers

of $\pi_k \geqslant 0$ along with the equality constraint $\mathbf{1}^T \boldsymbol{\pi} = 1$ to solve the problem. The optimal solution is given by:

$$\pi_k^* = \frac{\sum_{i=1}^N \gamma(x_{ik}) + \lambda_k - 1}{\sum_{k=0}^K \left(\sum_{i=1}^N \gamma(x_{ik}) + \lambda_k - 1 \right)}$$

Algorithm 1: EM Algorithm to estimate $\boldsymbol{\pi}$

Data: Given $\boldsymbol{z}$, $\boldsymbol{\sigma}^2$, $\sigma_0^2 = 0$, $\hat{\boldsymbol{s}}^2$;

1 Initialization: $\pi_k = 1/N$ for $k = 1, \ldots, K$, $\pi_0 = 1 - \sum_{k=1}^K \pi_k$;

2 Evaluate the incomplete log-likelihood $\ln p(\mathbf{z}|\boldsymbol{\pi}^{(0)})$;

3 **for** *r in 1:iteraitons* **do**

4   E step: Evaluate:

$$\gamma(x_{ik}) = \frac{\pi_k^{(r-1)} \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_k^2\right)}{\sum_{j=0}^K \pi_j^{(r-1)} \mathcal{N}\left(z_i; 0, \hat{s}_i^2 + \sigma_j^2\right)}, \quad k = 0, \ldots, K, \quad i = 1, \ldots, N$$

5   M step: update the parameter using:

$$\pi_k^{(r)} = \frac{\sum_{i=1}^N \gamma(x_{ik}) + \lambda_k - 1}{\sum_{k=0}^K \left(\sum_{i=1}^N \gamma(x_{ik}) + \lambda_k - 1 \right)}, \quad k = 0, \ldots, K$$

6   Evaluate the incomplete log likelihood $\ln p(\mathbf{z}|\boldsymbol{\pi}^{(r)})$;

7   Stop if either the parameter or the incomplete log-likelihood is convergent.

8 **end**

After obtaining the esimates for $\boldsymbol{\pi}$, denoted as $\hat{\boldsymbol{\pi}}$, we can substitute them back into (1), and combining with (2) to derive the posterior distribution of $\boldsymbol{\mu}$:

$p(\mu_i|z_i) \propto p(z_i|\mu_i)p(\mu_i)$

$$= \mathcal{N}\left(z_i; \mu_i, \hat{s}_i^2\right)\left(\hat{\pi}_0 \delta_0(\mu_i) + \sum_{k=1}^K \hat{\pi}_k \mathcal{N}\left(z_i; \mu_i; 0, \sigma_k^2\right)\right)$$

$$= \hat{\pi}_0 \delta_0(\mu_i) C_0 \exp\left\{ -\frac{(z_i - \mu_i)^2}{2\hat{s}_i^2} \right\} + \sum_{k=1}^K \hat{\pi}_k C_0 C_k \exp\left\{ -\frac{1}{2}\left(\frac{\sigma_k^2 + \hat{s}_i^2}{\sigma_k^2 \hat{s}_i^2} \mu_i^2 - 2\frac{z_i}{\hat{s}_i^2}\mu_i + \cdots \right) \right\} \quad (3)$$

$$\sim \pi_0 \delta_0(\mu_i) \mathcal{N}\left(\mu_i; z_i, \hat{s}_i^2\right) + \sum_{k=1}^K \pi_k \mathcal{N}\left(\mu_i; \frac{z_i \sigma_k^2}{\sigma_k^2 + \hat{s}_i^2}, \frac{\sigma_k^2 \hat{s}_i^2}{\sigma_k^2 + \hat{s}_i^2} \right)$$

Therefore, the posterior mean of μ_i is:

$$\mathbb{E}(\mu_i|z_i) = \sum_{k=0}^K \hat{\pi}_k \frac{z_i \sigma_k^2}{\sigma_k^2 + \hat{s}_i^2}, \quad \sigma_0^2 = 0, \quad i = 1, \ldots, N$$

1.3 Empirical Bayes for Spatially Structured Normal Mean under Homoskedasticity

Here, we consider the scenario where $\boldsymbol{\mu}$ is assumed to have spatial structure, which is completely different from the previous sparse case. The critical tool to handle this assumption is the Discrete Wavelet Transform.

Although I do not have any background knowledge about the Discrete Wavelet Transform and its associated properties, the information provided in this paper enables me to gain sufficient prior knowledge, which includes the following:

a) The sample Wavelet coefficients $\boldsymbol{w}$ of observed signals $\boldsymbol{z}$ after applying the Discrete Wavelet Transform can be expressed as:

$$\boldsymbol{w} = \mathbf{W}\boldsymbol{z}, \quad \mathbf{W} \in \mathbb{R}^{N \times N} \text{ is an orthogonal matrix}$$

b) **If the population mean of observed signals are spatially structured, then its corresponding Wavelet coeffcients are sparse**.

By incorporating these key insights, we can directly pull the problem back to the case of previous section (Empirical Bayes for Sparse Normal Mean). To see this, observe that under the assumption that observations $\boldsymbol{z}$, given a spatially structured $\boldsymbol{\mu}$, follow a isotropic Gaussian distribution:

$$z_i | \mu_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(z_i; \mu_i, s^2\right) \quad \Longrightarrow \quad \boldsymbol{z} | \boldsymbol{\mu} \sim \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}, s^2 \mathbf{I}_N\right)$$

Therefore, the sample Wavelet coefficients $\boldsymbol{w} = \boldsymbol{W}\boldsymbol{z}$ also follow a isotropic Gaussian distribution as well, since $\mathbf{W}$ is an orthogonal matrix (i.e., $\mathbf{W}\mathbf{W}^T = \mathbf{I}_N$):

$$\boldsymbol{w} | \tilde{\boldsymbol{\mu}} \sim \mathcal{N}\left(\boldsymbol{w}; \tilde{\boldsymbol{\mu}}, s^2 \mathbf{I}_N\right), \quad \tilde{\boldsymbol{\mu}} = \boldsymbol{W}\boldsymbol{\mu}$$

Most importantly, as mentioned before, $\tilde{\boldsymbol{\mu}}$ should be sparse. This is where the `ASH` method in the previous section becomes applicable.

Specifically, we can apply the `ASH` method to the sample Wavelet coefficients $\tilde{\boldsymbol{\mu}}$ to obtain an estimate of $\tilde{\boldsymbol{\mu}}$, denoted as $\hat{\tilde{\boldsymbol{\mu}}}$, and then perform the Inverse Discrete Wavelet Transform to obtain estimates of $\boldsymbol{\mu}$:

$$\hat{\boldsymbol{\mu}} = \mathbf{W}^T \hat{\tilde{\boldsymbol{\mu}}} \tag{4}$$

1.4 Empirical Bayes for Spatially Structured Normal Mean and Variance under Heteroscedasticity

This section constitutes the major contribution of this paper, as it generalizes the assumption of homoskedasticity from the previous section to the case of heteroscedasticity and develops a method for estimating both the mean and variance, referred to as `SMASH` [2]. Here, we assume:

$$z_i | \mu_i \overset{\text{indep}}{\sim} \mathcal{N}(z_i; \mu_i, s_i^2) \quad \Longrightarrow \quad \boldsymbol{z} | \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \mathbf{S}), \ \mathbf{S} = \text{diag}(s_1^2, \ldots, s_N^2) \tag{5}$$

where the $\boldsymbol{\mu}$ is still assumed to have spatial structure. Additionally, $\boldsymbol{s}^2 = (s_1^2, \ldots, s_N^2)$ is also assumed to have spatial structure.

Then, similar to the previous section, we can first apply the Discrete Wavelet Transform to the observed signal $\boldsymbol{z}$ to obtain the sample Wavelet coefficients $\boldsymbol{w} = \boldsymbol{W}\boldsymbol{z}$. Let us denote $\tilde{\boldsymbol{\mu}} = \boldsymbol{W}\boldsymbol{\mu}$, which should be sparse. The Wavelet coefficients $\boldsymbol{w}$ will then follow a Multivariate Normal distribution given by:

$$\boldsymbol{w}|\tilde{\boldsymbol{\mu}} \sim \mathcal{N}\big(\boldsymbol{w}; \tilde{\boldsymbol{\mu}}, \mathbf{WSW}^T\big) \tag{6}$$

To estimate $\boldsymbol{\mu}$, suppose $\boldsymbol{s}^2$ is given. Interestingly, when comparing (6) with (2), we observe that they are almost in the same format, except that $\boldsymbol{S}$ in (2) is a diagonal matrix, whereas $\mathbf{WSW}^T$ in (6) is not necessarily diagonal.

The approach taken in this paper to address this issue is straightforward: simply ignore the off-diagonal terms of $\mathbf{WSW}^T$. The authors argue that many existing literature have adopted similar strategies. By doing this, we have:

$$\boldsymbol{w}|\tilde{\boldsymbol{\mu}} \sim \mathcal{N}\big(\boldsymbol{w}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{S}}\big), \quad \tilde{\boldsymbol{S}} = \operatorname{diag}(\tilde{\boldsymbol{s}}^2), \; \tilde{\boldsymbol{s}}^2 = \operatorname{diag}\big(\mathbf{WSW}^T\big) \tag{7}$$

Therefore, (7) and (2) are in exactly the same format. The problem is then simplified to applying ASH on $\boldsymbol{w}$ based on assumption (7).

Next, to estimate $\boldsymbol{s}^2$ with given $\boldsymbol{\mu}$, let us denote $\tilde{z}_i^2 = (z_i - \mu_i)^2$. By (5), it follows that:

$$\tilde{z}_i^2|s_i^2 \overset{\text{indep}}{\sim} s_i^2 \chi_1^2, \quad i = 1, \dots, N \tag{8}$$

In this paper, the right-hand side Chi-Squared related distribution is approximated by a Normal distribution, resulting in:

$$\tilde{z}_i^2|s_i^2 \overset{\text{indep}}{\sim} \mathcal{N}\Big(\tilde{z}_i^2; s_i^2, \frac{2}{3}\tilde{z}_i^4\Big) \implies \tilde{\boldsymbol{z}}^2|\boldsymbol{s}^2 \sim \mathcal{N}\big(\tilde{\boldsymbol{z}}^2; \boldsymbol{s}^2, \boldsymbol{\Psi}\big), \; \boldsymbol{\Psi} = \frac{2}{3}\operatorname{diag}(\tilde{z}_1^4, \dots, \tilde{z}_N^4) \tag{9}$$

where the term $\frac{2}{3}\tilde{z}_i^4$ arises from the fact that it is an unbiased estimator of $\operatorname{Var}(\tilde{z}_i^2|s_i^2)$ under the (8).

Then, recalling the assumption that $\boldsymbol{s}^2$ exhibits spatial structure, we observe that (9) and (5) are exactly in the same format. Therefore, we can apply the Discrete Wavelet Transform to the observed signal $\tilde{\boldsymbol{z}}^2$ to obtain the sample Wavelet coefficients $\tilde{\boldsymbol{w}}^2 = \mathbf{W}\tilde{\boldsymbol{z}}^2$. Let us denote $\tilde{\boldsymbol{s}}^2 = \mathbf{W}\boldsymbol{s}^2$, which should be sparse. The Wavelet coefficients $\tilde{\boldsymbol{w}}^2$ will then follow a Multivariate Normal distribution given by:

$$\tilde{\boldsymbol{w}}^2|\tilde{\boldsymbol{s}}^2 \sim \mathcal{N}\big(\tilde{\boldsymbol{w}}^2; \tilde{\boldsymbol{s}}^2, \mathbf{W}\boldsymbol{\Psi}\mathbf{W}^T\big) \tag{10}$$

We can see that (10) and (6) are exactly in the same format, and the problem is simplied to first ignore the off-diagonal part of $\mathbf{W}\boldsymbol{\Psi}\mathbf{W}^T$, resulting in:

$$\tilde{\boldsymbol{w}}^2|\tilde{\boldsymbol{s}}^2 \sim \mathcal{N}\big(\tilde{\boldsymbol{w}}^2; \tilde{\boldsymbol{s}}^2, \tilde{\boldsymbol{\Psi}}\big), \quad \tilde{\boldsymbol{\Psi}} = \operatorname{diag}(\tilde{\boldsymbol{\psi}}^2), \; \tilde{\boldsymbol{\psi}}^2 = \operatorname{diag}\big(\mathbf{W}\boldsymbol{\Psi}\mathbf{W}^T\big) \tag{11}$$

and then apply ASH on $\tilde{\boldsymbol{w}}^2$ based on assumption stated in (11).

Finally, to estimate both $\boldsymbol{\mu}$ and $\boldsymbol{s}^2$, this paper proposed an iterative algorithm:

Algorithm 2: SMASH

    Data: Given observations $\boldsymbol{z} = (z_1, \ldots, z_N)$;

1 Initialization: $\boldsymbol{s}_{(1)}^2 = \left(s_{1,(1)}^2, \ldots, s_{N,(1)}^2\right)$, such that:

$$s_{i,(1)}^2 = \frac{1}{2}\left((z_i - z_{t-1})^2 + (z_i - z_{i+1})^2\right), \quad z_{N+1} = z_1, \ z_0 = z_N$$

2 Apply Discrte Wavelet Transform on $\boldsymbol{z}$: $\ \boldsymbol{w} = \boldsymbol{W}\boldsymbol{z}$;

3 for *r in 1:iteraitons* **do**

4   Esitmate $\hat{\tilde{\boldsymbol{\mu}}}_{(r)}$ under assumption (7): apply `ASH` on $\boldsymbol{w}$, with variance of each coordinate w_i given by:

$$\tilde{\boldsymbol{s}}^2 = \mathrm{diag}\left(\mathbf{W}\mathrm{diag}(\boldsymbol{s}_{(r)}^2)\mathbf{W}^T\right)$$

5   Esitmate $\hat{\boldsymbol{\mu}}_{(r)}$ by Inverse Discrete Wavelet Transform (4):

$$\hat{\boldsymbol{\mu}}_{(r)} = \mathbf{W}^T\hat{\tilde{\boldsymbol{\mu}}}_{(r)}$$

6   Apply Discrte Wavelet Transform on $\tilde{\boldsymbol{z}}^2 = \left(\boldsymbol{z} - \hat{\boldsymbol{\mu}}_{(r)}\right)^2$: $\ \tilde{\mathbf{w}}^2 = \mathbf{W}\tilde{\boldsymbol{z}}^2$;

7   Esitmate $\hat{\tilde{\boldsymbol{s}}}_{(r)}^2$ under assumption (11): apply `ASH` on $\tilde{\boldsymbol{w}}^2$, with variance of each coordinate w_i given by:

$$\tilde{\boldsymbol{\psi}}^2 = \mathrm{diag}\left(\mathbf{W}\boldsymbol{\Psi}\mathbf{W}^T\right), \ \ \boldsymbol{\Psi} = \frac{2}{3}\mathrm{diag}(\tilde{z}_1^4, \ldots, \tilde{z}_N^4)$$

8   Esitmate $\hat{\boldsymbol{s}}_{(r)}^2$ by Inverse Discrete Wavelet Transform (4):

$$\hat{\boldsymbol{s}}_{(r)}^2 = \mathbf{W}^T\hat{\tilde{\boldsymbol{s}}}_{(r)}^2$$

9   Stop if the parameters are convergent.

10 end

---

2 Result and Discussion

The codes and results of this section are available at Github.

2.1 Clarification

For replicating the key results of this paper, firstly I would like to clarify which parts I implement by myself, and which parts are based on the information presented in the paper.

    The two **core algorithms**: `SMASH` method [2] in this paper, and `ASH` method [1] in another paper, are implemented by myself. Essentially, the `SMASH` method applies the `ASH` method to the Wavelet coefficients, and both algorithms are based on the Empirical Bayes framework.

For generating simulation data and performing the Wavelet Transform, I refer to the code provided in the paper and existing packages.

a) **Generating Simulation Data**: In Appendix C of this paper, they provided a list of 7 mean functions and 5 variance functions, correspond to different types of structured mean and variance, for generating Gaussian data. To generate the simulation data, I utilized the corresponding functions they provided in their GitHub repository. As this aspect of the study is unrelated to the core algorithm presented in the paper, I believe it is acceptable to leverage their code for this purpose.

b) **Wavelet Transform**: In this paper, the "non-decimated" Wavelet Transform is chosen, with the argument that it is a standard technique in signal processing known to enhance performance. This part is corresponding to another area (i.e. signal processing), and I do not have any background knowledge about it. I can understand the high-level concepts about why should we use it, but implementing it is challenging for me. Therefore, I use the function "wd" in R package `wavethresh` to do the "non-decimated" Wavelet Transform. Note that this paper also use this package to compute the "non-decimated" Wavelet Transform.

2.2   Results

The first result I present is my own replication of the `ASH` method [1], comparing it to the `ash` function in R package `ashr` developed by the author. This comparison is conducted on simulation data and is crucial to validate the correctness of my replication of `ASH`, as it serves as a crucial component of the `SMASH` method [2] introduced in this paper.

To perform the comparison, I generate simulation data z using the "Spike" mean function, which represents one type of spatially structured mean. I then apply a "non-decimated" Wavelet Transform to z to obtain a group of Wavelet coefficient $\{w_g\}_{g=1,...,G}$. Next, I apply my own replication of the `ASH` method and also utilize the `ash` function from the `ashr` package on each single w_g, and the summary statistics for important quantities, such as the estimated $\hat{\pi}$, incomplete likelihood, MSE of the estimated posterior mean compared to the true mean, etc., are reported. It is worth noting that the parameters used in the `ash` function are set as `mixcompdist='normal'` and `optmethod='mixEM'`, which correspond to a mixture Normal prior distribution and the EM optimization method, respectively, to ensure comparability. The results are shown in Figure 1, and the discussions of this result are provided in the subsequent subsection.

Next, comes to the results of my own replication of the `SMASH` method [2]. The procudure of data generation and "non-decimated" Wavelet Transform are the same with previous result. I compare the results of following 3 algorithms:

a) Applying algorithm 2 using my own replication of `ASH` method.

b) Applying algorithm 2 using `ash` function in `ashr` package.

c) Directly using `smash` function in `smashr` package.

The MSE of the estimated posterior mean compared to the true mean of these algorithms are reported, and the resuls are shown in Figure 2, and the discussions of this result are provided in the subsequent subsection.
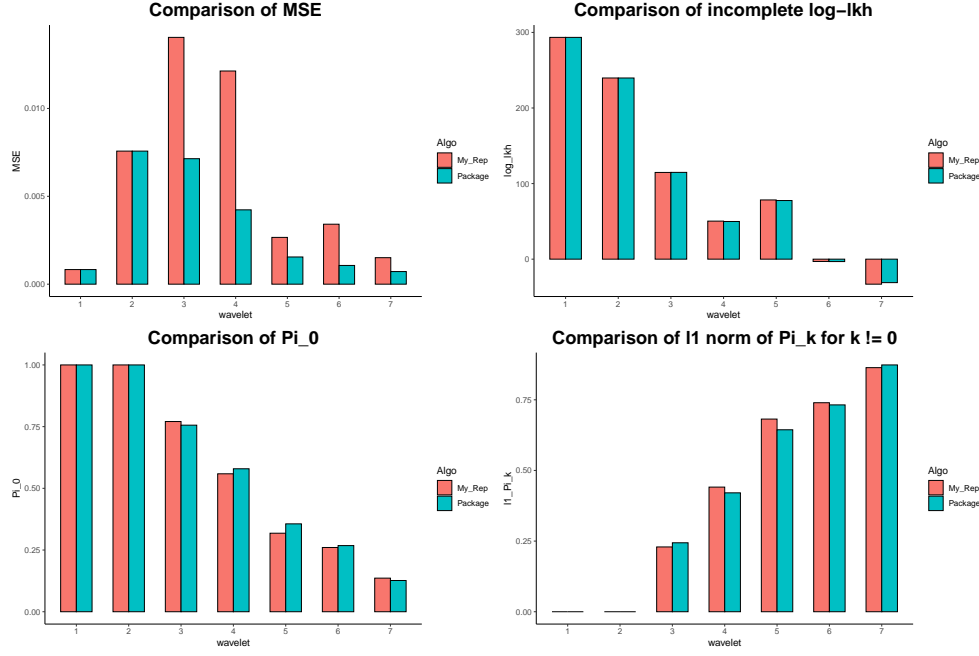
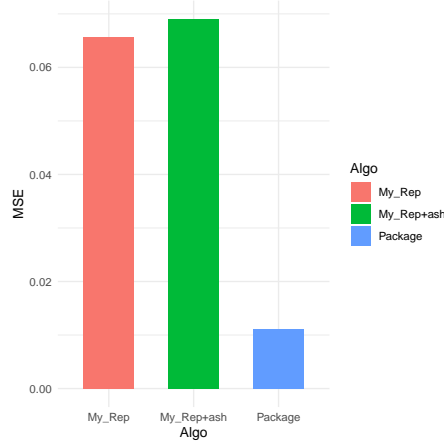Figure 1: Comparison of ASH Replication with `ash` function in Package

Figure 2: Comparison of SMASH Replication with its Alternatives

2.3 Discussion

The Figure 1 presents a comparison for my replication of `ASH` method with the `ash` function based on one realization of simulated dataset with 2^7 data points. I also conducted checks on other realizations by setting different seeds. Notably, when π_0 is very close to 1, the outcomes of my replication are exactly the same as those produced by the `ash` function. However, when π_0 deviates from 1, the MSE of my replication tends to be larger than that of the standard package. Nevertheless, the other quantities remain highly similar, indicating that my replication of `ash` is nearly correct.

The issue with my replication for this part lies in the slow convergence of the standard EM algorithm. It is important to note that the `ash` function in the `ashr` package utilizes an accelerated EM algorithm from the `R` package `SQUAREM`. Moreover, the advanced EM algorithm has the potential to achieve more accurate estimation.

The Figure 2 presents a comparison of my replication of the `SMASH` method with other alternatives mentioned in the previous subsection, based on the same simulation dataset as before. However, I encountered difficulties in replicating the `SMASH` method, which could be attributed to several reasons:

a) The Inverse Wavelet Transform given by (4) may not be numerically stable. Specifically, I found that the matrix $\mathbf{W}\mathbf{W}^T$ or $\mathbf{W}^{-1}\mathbf{W}$ deviates significantly from the identity matrix, with off-diagonal terms that are considerably larger than zero.

b) In my numerical experiment, the distribution approximation (9) proposed by the original paper can result in estimated variances $\hat{\boldsymbol{s}}^2$ being negative. This is because the domain of the Normal distribution does not necessarily exclude non-negative values. Although Appendix A of the paper provides a detailed procedure for estimating the variance, they did not mention this issue. To address this problem, I handled negative outcomes by setting them to a very small number (e.g., $1e-4$).

c) There may be additional tasks that need to be performed. For example, in Appendix B.3 of the paper, they imported `MATLAB` code developed by others into `R` to perform circulant shifts of the data.

3 Conclusion

This paper presents an intuitive and effective method within the Empirical Bayes framework to address the issue of homoskedasticity when estimating spatially structured mean and variance.

The Discrete Wavelet Transform establishes a connection between the cases of sparsity and spatial structure through the property: "if the population mean of observed signals exhibits spatial structure, then its corresponding Wavelet coefficients will be sparse."

Consequently, the estimation of spatially structured mean reduces to applying Empirical Bayes techniques for sparse mean estimation on Wavelet coefficients. Similarly, the estimation of spatially structured variance involves applying Empirical Bayes techniques for sparse mean estimation on Wavelet coefficients of square of demeaned observations, while making certain distributional approximations.

References

[1] M. Stephens, "False discovery rates: a new deal," *Biostatistics*, 2017.

[2] M. S. Zhengrong Xing, Peter Carbonetto, "Flexible signal denoising via flexible empirical bayes shrinkage," *Journal of Machine Learning Research*, 2021.