Sankhyā A : The Indian Journal of Statistics
https://doi.org/10.1007/s13171-019-00177-0

Bayesian Fusion Estimation via t Shrinkage

Qifan Song and Guang Cheng
Purdue University, West Lafayette,, USA

Abstract

Shrinkage prior has gained great successes in many data analysis, however, its applications mostly focus on the Bayesian modeling of sparse parameters. In this work, we will apply Bayesian shrinkage to model high dimensional parameter that possesses an unknown blocking structure. We propose to impose heavy-tail shrinkage prior, e.g., t prior, on the differences of successive parameter entries, and such a fusion prior will shrink successive differences towards zero and hence induce posterior blocking. Comparing to conventional Bayesian fused LASSO which implements Laplace fusion prior, t fusion prior induces stronger shrinkage effect and enjoys a nice posterior consistency property. Simulation studies and real data analyses show that t fusion has superior performance to the frequentist fusion estimator and Bayesian Laplace fusion prior. This t fusion strategy is further developed to conduct a Bayesian clustering analysis, and our simulations show that the proposed algorithm compares favorably to classical Dirichlet process modeling.

AMS (2000) *subject classification.* Primary 62F15; Secondary 62J07.
Keywords and phrases. t shrinkage prior, Bayesian fusion, Bayesian clustering, Posterior consistency

1 Introduction

High dimensionality plays an important role in modern statistical applications such as genomics, image processing, finance and etc. An overview for the development of high dimensional analysis can be found in van der Geer and Bühlmann (2011) and references therein. To overcome ill-posed problems that involve high dimensional parameters, one usually assumes that the true parameter value lies in a low dimensional subspace. To obtain such low dimensional estimation, the idea of regularization is commonly used, via penalized likelihood approaches or using informative prior specifications. Various penalty functions have been proposed for consistent frequentist estimation, including LASSO (Tibshirani, 1996), Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001), adaptive LASSO (Zou, 2006) and

Minimax Concave Penalty (MCP) (Zhang, 2010). For high dimensional Bayesian inferences, sparsity induced priors, such as spike-and-slab prior (Jiang, 2007; Liang et al., 2013; Narisetty and He, 2014; Song and Liang, 2014; Yang et al., 2015; Castillo et al., 2015; Scott and Berger, 2010; Johnson and Rossel, 2012), are widely used for model selection. Fused LASSO (Tibshirani et al., 2005) considers another type of low dimensional embedding of a high dimensional parameter $\theta = (\theta_i)_{i=1}^p$ where the successive differences $\vartheta_i = \theta_i - \theta_{i-1}$ are assumed to be sparse as well, in other words, there exists a consecutive block partition of θ_i's, such that θ_i's are constant within each block. Fused LASSO method proposes a penalty function $\lambda_1 \sum |\theta_i| + \lambda_2 \sum |\vartheta_i|$ which consists of two terms that encourage sparsity among θ_i's and ϑ_i's respectively.

In this work, we consider the following Gaussian mean problem:

$$y_i = \theta_i^* + \varepsilon_i \tag{1.1}$$

where ε_i's are iid normal error with unknown variance σ^2. Similarly to fused LASSO applications, we also assume true parameter θ^* is blocky in the sense that there exists a partition $\{\mathcal{B}_1^*, \ldots, \mathcal{B}_s^*\}$ of $\{1, ..., n\}$ such that θ_i^*'s are constant for all $i \in \mathcal{B}_k^*$. Correspondingly, we define set $G^* = \{2 \leq i \leq n : \vartheta_i^* := \theta_i^* - \theta_{i-1}^* \neq 0\}$ whose number of elements is supposed to be much smaller than n. We are interested in conducting Bayesian structure recovery of θ^*. Motivated by the L_1 fusion penalty used by fused LASSO estimator (Tibshirani et al., 2005), as well as the development of the Bayesian LASSO (Park and Casella, 2008), Kyung et al. (2010) introduced Bayesian fused LASSO by imposing independent Laplace priors on all successive differences. The implementation of Laplace shrinkage prior can significantly reduce the posterior sampling costs compared to spike-and-slab modeling, and conceptually, the Laplace prior can be nicely interpreted as a Bayesian counterpart of L_1 penalty. However, many recent Bayesian theoretical developments show that in the context of sparse linear regression models, Laplace prior fails to achieve satisfactory posterior contraction (Castillo et al., 2015; Bhattacharya et al., 2015; Song and Liang, 2017). It is believed that the posterior inconsistency of Laplace prior is due to its exponentially light tail, and Song and Liang (2017) suggests to use heavy tail prior distribution for sparse linear regression models, which can induce sufficient Bayesian shrinkage effect and thereafter guarantee to recover the sparsity structure.

We find that the above phenomenon holds for Bayesian fusion estimation as well: imposing Laplace prior on $(\theta_i - \theta_{i-1})$ leads to a smoothly varying θ_i estimation rather than a blocky θ, thus it fails to identify the

blocking structure. Therefore, in this paper, we propose to use independent student-t priors on successive differences $\theta_i - \theta_{i-1}$ for a Bayesian fusion problem. Our results show that such a simple t fusion Bayesian modeling leads to very accurate posterior estimation. More importantly, comparing with Laplace prior or frequentist L_1 penalization, its performance on detecting the blocking structure is much better. The asymptotic posterior convergence induced by t fusion prior is investigated as well. A related Bayesian work is Shimamura et al. (2018) who proposed to use a Normal-Exponential-Gamma (NEG) prior for the successive differences. However, their Bayesian inference is only based on the maximum a posterior (MAP) estimator, while our application tries to fully utilize the whole posterior distributional information.

Furthermore, we consider a practically useful extension to Bayesian fusion estimation. Instead of assuming that θ^* has a *consecutive* blocking structure, it is more realistic to assume that θ^* possesses an unknown clustering structure. In other words, θ_i^*'s that are not necessarily consecutive can still share the same value. In a broader scope, such a clustering problem can be viewed as a simplest example of subgroup analysis where we assume that a subject-related parameter θ follows an unknown grouping structure. For example, in clinical trial studies, the treatment effects may vary across different subpopulations, but remain the same for the patients belonging to the same subpopulation. If one can correctly identify the subpopulation structure, then specific medical therapies can be prescribed for each subpopulation to maximize the treatment effectiveness. The existing models for Bayesian cluster analysis (Wade and Ghahramani, 2018; Heller and Ghahramani, 2005; Mozeika and Coolen, 2018; Berger et al., 2014) usually impose discrete priors on the clustering structure, along with a conditional prior on θ given specific clustering structure. In contrast, we propose to directly model the parameter θ via t fusion prior.

This paper is organized as follows. In Section 2, we study the Bayesian fusion problem with t prior specification. We will present the posterior asymptotic result, and discuss its difference from the Laplace prior. In Section 3, we will use the t fusion prior to solve the clustering problem. Several simulation studies and one real data application are presented in Section 4. Finally, Section 5 provides more discussions and remarks. All technical proofs are postponed to the Appendix.

Throughout this work, the following notations are used. Given two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \succ b_n$ means $\lim(a_n/b_n) = \infty$ and $a_n \asymp b_n$ means $-\infty < \liminf(a_n/b_n) \leq \limsup(a_n/b_n) < \infty$. $\|x\|$ and $\|x\|_1$ denote L_2 and L_1 norms of vector x.

2 Bayesian Fusion Via t Shrinkage Prior

2.1. Bayesian Modeling Suppose we observe independent data $\{y_i : i = 1, \ldots, n\}$ following model (1.1). The indexing of the data has certain practical or scientific meaning, under which we can assume that the parameter vector θ^* is "stepwise", in the sense that most of the successive differences $\vartheta_i = \theta_i - \theta_{i-1}$ are exactly 0. To induce the sparsity for both θ_i's and ϑ_i's, (Tibshirani et al., 2005) proposed the following fused LASSO estimator

$$\widehat{\theta}^{\mathrm{FL}} = \arg\min \left(\frac{\|y-\theta\|^2}{2} + \lambda_1 \sum_{i=1}^{n} |\theta_i| + \lambda_2 \sum_{i=2}^{n} |\vartheta_i| \right).$$

If one is not interested in pursuing the sparsity of θ's, then a fusion estimator (Rinaldo et al., 2009) can be used

$$\widehat{\theta}^{\mathrm{F}} = \arg\min \left(\frac{\|y-\theta\|^2}{2} + \lambda \sum_{i=2}^{n} |\vartheta_i| \right) = \arg\min \left(\frac{\|y-\theta\|^2}{2} + \lambda \sum_{i=2}^{n} |\theta_i - \theta_{i-1}| \right), \tag{2.1}$$

for some tuning parameter λ. The above objective functions are both convex and fast computation algorithms are developed, e.g. Liu et al. (2010) & Tibshirani et al. (2005). The penalty term $\lambda \sum_{i=2}^{n} |\theta_i - \theta_{i-1}|$ can be interpreted as the negative logarithm of prior density used for Bayesian inferences, therefore, a natural Bayesian expansion to Eq. 2.1 is Laplace (double exponential) prior modeling (Kyung et al., 2010; Shimamura et al., 2018). To account for the unknown variance parameters σ^2 and θ_1, a convenient prior specification could be

$$\begin{array}{ll} \sigma^2 & \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma), \quad \theta_1|\sigma^2 \sim N(0, \sigma^2\lambda_1), \\ (\theta_i - \theta_{i-1})|\sigma^2 & \sim \text{Laplace}(\lambda/\sigma), \text{ for all } i = 2, \ldots, n \end{array} \tag{2.2}$$

where Laplace(a) denotes the distribution with cdf $f(x) \propto \exp(-a|x|)$. Note that besides the fusion prior imposed on consecutive differences, we also assign a prior on θ_1, this ensures the joint prior of θ is a proper prior distribution.

According to Andrews and Mallows (1974), the above Laplace(λ/σ) prior can be rewritten as a scale mixture of normal distributions:

$$(\theta_i - \theta_{i-1})|\sigma^2, \lambda_i \sim N(0, \lambda_i\sigma^2), \quad \lambda_i \sim \exp(-\lambda^2/2),$$

where $\exp(a)$ denotes the exponential distribution $f(x) \propto \exp(-ax)$. This hierarchical representation for Laplace prior leads to a Gibbs

sampling update that is similar to the Bayesian LASSO (Park and Casella, 2008):

$$\begin{aligned}
\lambda_i^{-1} \sim\ & \text{Inverse-Gaussian}(\lambda\sigma/|\theta_i - \theta_{i-1}|, \lambda^2) \quad \text{for all } i = 2, \ldots, n, \\
\sigma^2 \sim\ & \text{Inverse-Gamma}\left(a_\sigma + n, b_\sigma + \tfrac{\|y-\theta\|^2}{2} + \tfrac{\theta_1^2}{2\lambda_1} + \sum_{i=2}^{n} \tfrac{(\theta_i - \theta_{i-1})^2}{2\lambda_i}\right), \\
\theta_i \sim\ & N(\mu_i, \nu_i) \quad \text{for } i = 1, \ldots, n
\end{aligned} \tag{2.3}$$

where $\nu_i^{-1} = 1/\sigma^2 + 1/\lambda_{i+1}\sigma^2 + 1/\lambda_i\sigma^2$, $\mu_i = \nu_i(y_i/\sigma^2 + \theta_{i+1}/\lambda_{i+1}\sigma^2 + \theta_{i-1}/\lambda_i\sigma^2)$ for $i = 1, \ldots, n$; λ_{n+1} is considered to be infinite, and θ_0 is consider to be 0; Inverse-Gaussian(a, b) denotes inverse Gaussian distribution with cdf $f(x) \propto x^{-3/2} \exp[-b(x-a)^2/(2a^2x)]$.

Despite the popularity of Laplace prior in many applications, recent Bayesian works (Castillo et al., 2015; Bhattacharya et al., 2015; Song and Liang, 2017) point out that, if we impose independent $\theta_i|\sigma^2 \sim \text{Laplace}(\lambda/\sigma)$ for linear regression models with a sparse regression coefficient θ, then the induced posterior has only a sub-optimal contraction rate, or even diverges. In other words, the posterior distribution of θ doesn't contract into a small neighborhood around true value θ^* appropriately. For a blocky parameter θ, we observe similar empirical results, as showed in the toy example in Section 2.3: the Laplace fusion prior fails to shrink the observations, which belongs to the same block, towards a same value. Hence, the resultant Bayesian estimate of θ doesn't have a step-wise pattern at all.

Following the theoretical discovery of Song and Liang (2017), we consider using a class of heavy tailed priors for the successive differences ϑ_i's. Specifically, this work will assign a t shrinkage prior:

$$\begin{aligned}
\sigma^2 \quad & \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma), \quad \theta_1|\sigma^2 \sim N(0, \sigma^2\lambda_1), \\
(\theta_i - \theta_{i-1})|\sigma^2 \quad & \sim t_{df}(s\sigma), \quad \text{for all } i = 2, \ldots, n
\end{aligned} \tag{2.4}$$

where $t_a(b)$ denotes t distribution with degree of freedom a and scale parameter b. Note that the above t distribution can be rewritten as an inverse-gamma scaled Gaussian mixture as

$$(\theta_i - \theta_{i-1})|\sigma^2, \lambda_i \sim N(0, \lambda_i\sigma^2), \quad \lambda_i \sim \text{Inverse-Gamma}(a_t, b_t),$$

where a_t, b_t satisfy $df = 2a_t$ and $s = \sqrt{b_t/a_t}$. Under this t prior, the posterior distribution still allows a full conditional Gibbs sampler, where the updates for θ_i's and σ^2 are exactly the same as in Eq. 2.3 and the update of λ_i's follows

$$\lambda_i \sim \text{Inverse-Gamma}\left(a_t + 1/2, b_t + \frac{(\theta_i - \theta_{i-1})^2}{2\sigma^2}\right) \quad \text{for all } i = 2, \ldots, n.$$

To further understand the difference between the Laplace fusion prior and t fusion prior, we compare their conditional prior $\pi(\theta_i|\theta_{i-1}, \theta_{i+1}, \sigma)$ for $1 < i < n$. Note that this conditional distribution is completely determined by the fusion priors imposed on $\theta_{i+1} - \theta_i$ and $\theta_i - \theta_{i-1}$. Figure 1 plots the function $-\log[\pi(\theta_i|\theta_{i-1} = -1, \theta_{i+1} = 1, \sigma = 1)]$, up to a constant, for both prior specifications. It is clear that the conditional t fusion prior allocates most of its prior mass at the two small neighborhoods centered at θ_{i-1} and θ_{i+1}, given a sufficiently small scale parameter s. In other words,

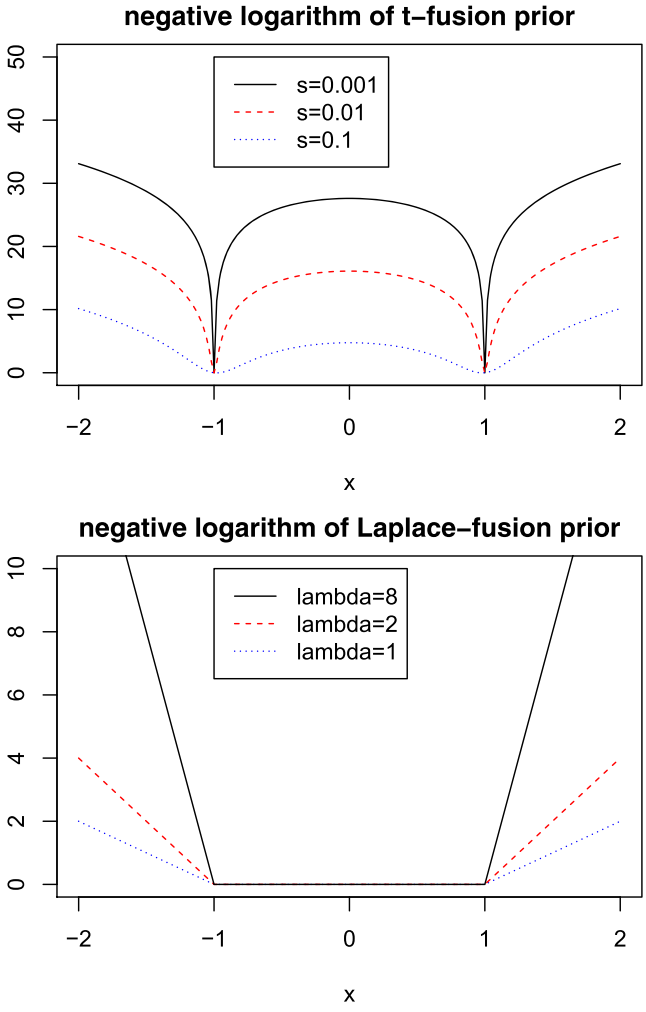

Figure 1: The negative logarithm of conditional prior $-\log[\pi(\theta_i|\theta_{i-1} = -1, \theta_{i+1} = 1, \sigma = 1)]$ under different hyperparameter values

the prior introduces a strong shrinkage effect on θ_i, towards either θ_{i-1} or θ_{i+1}. Therefore, for all $i = 2, \ldots, n-1$, θ_i will merge with either θ_{i-1} or θ_{i+1} in the posterior distribution, which thereafter induces a posterior blocking structure. On the other hand, the conditional Laplace fusion prior has a uniform prior density within the interval $[\theta_{i-1}, \theta_{i+1}]$. Hence, it doesn't encourage the posterior of θ_i to be grouped with either θ_{i-1} or θ_{i+1}. It is worth to mention that the NEG fusion prior (Shimamura et al., 2018) also has a similar pattern for its conditional prior density function. But a critical difference between t prior and NEG prior, is that the density for NEG prior is non-differentiable at 0. Therefore, the MAP of NEG fusion prior possesses an exact blocking structure, and Shimamura et al. (2018) only use this MAP for Bayesian inferences rather than the whole posterior distributional information. But the t prior is continuously differentiable everywhere (the functions displayed in the upper plot of Fig. 1 are actually smooth at -1 and 1). Hence, its MAP doesn't have blocky structure, and in this work we will utilize all the posterior samples for the Bayesian analysis.

2.2. Posterior Contraction of Bayesian t Fusion In this section, we study the theoretical performance of t fusion prior specification (2.4). Our theoretical investigation follows the framework of Song and Liang (2017), which studies the posterior convergence rate of coefficient β in high dimensional sparse regression models $y = X\beta + \varepsilon$. Note the model (1.1) can also be represented as a sparse linear regression, where the design matrix X is a n by n matrix whose lower triangle entries are all 1 and $\beta = (\theta_1, \vartheta_2, \ldots, \vartheta_n)$ is a unknown sparse vector. The following theorem studies the general posterior convergence properties given an independent prior over all ϑ_i's.

Theorem 2.1 (Posterior consistency)**.** *Assume that* $|G^*| \prec n/\log n$*, and the prior specification follows that* $\sigma^2 \sim$ *Inverse-Gamma*(a_σ, b_σ)*,* θ_1 *and* ϑ_i*'s are conditionally independent given* σ^2 *with prior density* $\pi(\theta_1, \vartheta_2, \ldots, \vartheta_n|\sigma) \propto (1/\sigma)^n f_\theta(\theta_1/\sigma) \prod_{i=2}^n f_\vartheta(\vartheta_i/\sigma)$*. Furthermore, if*

$$\begin{aligned}
&\int_{-|G^*|\log n/n^2}^{|G^*|\log n/n^2} f_\vartheta(x)dx \geq 1 - n^{-(1+u)}, \text{ for some } u > 0,\\
&-\log(\underline{\pi}_\vartheta) = O(\log n), \quad \text{where } \underline{\pi}_\vartheta = \min_{|x|\leq \max_i |\vartheta_i^*/\sigma^*|+1} f_\vartheta(x),\\
&-\log(\underline{\pi}_\theta) = O(|G^*|\log n), \quad \text{where } \underline{\pi}_\theta = \min_{|x|\leq |\theta_1^*/\sigma^*|+1} f_\theta(x),\\
&a_\sigma \log(1/b_\sigma) + b_\sigma/\sigma^{*2} + (a_\sigma + 2)\log(\sigma^{*2}) = O(|G^*|\log n),
\end{aligned} \tag{2.5}$$

then there exist a constant M *and* $\epsilon_n \asymp \sqrt{|G^*| \log n/n}$*, such that the posterior distribution satisfies*

$$\pi(\|\theta - \theta^*\| \geq M\sigma^* \sqrt{n}\epsilon_n | y) \to 0,$$

where the convergence holds in probability or in L_1 *w.r.t. the probability measure of* y*.*

The proof closely follows the Theorem A.1 in Song and Liang (2017), and for the sake of readability, the proof is provided in the Appendix. The first inequality of sufficient condition set (2.5) requires that the prior imposed on ϑ_i's is highly concentrated around zero, such that it induces sufficient Bayesian shrinkage effect for those ϑ_i's whose true values are 0. The rest inequalities of Eq. 2.5 essentially require that the prior density at true parameter is at least of order $e^{-cn\epsilon_n^2}$ for some c, and this helps to prevent over-shrinkage for those ϑ_i's whose true values are not 0. Similar conditions, which need the prior to be "thick" at true parameter values, are regularly used in Bayesian literature (Jiang 2007; Kleijn and van der Vaart 2006b; Ghosal et al., 2000, 2007). Given the concrete forms for prior density f_θ and f_ϑ, the second and third inequalities of Eq. 2.5 are equivalent to some upper bound constraints on the magnitude of θ_1^* and $\max_i |\vartheta_i^*|$ (see e.g., Corollary 2.1). The fourth inequality of Eq. 2.5 trivially holds for any fixed a_σ and b_σ if σ^{*2} is assumed to be a constant. If the unknown error variance is supposed to be varying with respect to n, e.g., the studies of Gaussian sequence models (Johnstone, 2010) commonly assume that $\sigma^{*2} \propto n^{-1}$, then one can choose a fixed a_σ and $b_\sigma \asymp n^{-\kappa}$ for some $\kappa > 0$. Under such a choice, condition (2.5) holds as long as $-\kappa \log n \leq \log(\sigma^{*2}) < K \log n$ for some positive constant K.

The above result states that almost all the posterior mass contracts into a neighborhood of θ^* with radius $M\sigma^*\sqrt{n}\epsilon_n$, that is, the posterior convergence rate is of order $\sigma^*\sqrt{|G^*| \log n}$. Note that if the partition index set G^* were known, the oracle rate of contraction turns out to be $O(\sigma^*\sqrt{|G^*|})$. Hence, the Bayesian shrinkage achieves the ideal risk up to a logarithmic term in n. In frequentist literature, Theorem 2.7 of Rinaldo et al. (2009) showed that the convergence rate of fused LASSO is no larger than $O(\sigma^*\sqrt{|G^*| \log |G^*|})$. However, this rate is not directly comparable to ours, since an additional minimal signal strength condition, which ensures that G^* can be fully recovered in probability, is imposed in Rinaldo et al. (2009).

The posterior distribution of ϑ is always continuous, and doesn't directly provide Bayesian inferences for the block structure, or equivalently, the unknown G^*. The following result characterizes some asymptotic performance of posterior block partition via discretization.

Theorem 2.2 (Posterior selection)**.** *Assume the conditions of Theorem 2.1 hold, denote* $G(\theta, \sigma) = \{i : |\vartheta_i/\sigma| < \epsilon_n/n\}$*, then the posterior of* $G(\theta, \sigma)$ *satisfies*

$$\pi(\{|G(\theta, \sigma) \backslash G^*| > \delta |G^*|\}|y) \to 0,$$

for some fixed constant δ*, where the convergence holds in probability and in* L_1*.*

Therefore, the posterior distribution of $G(\theta, \sigma)$, which is induced by the posterior of (θ, σ^2) and mapping $G(\cdot, \cdot)$, can be treated as a discrete posterior distribution for the unknown G^* and used for Bayesian block partition selection. Theorem 2.2 essentially claims that the number of false positive selections via discretization $G(\cdot, \cdot)$ is bounded in posterior probability. Such a result is comparable to the model selection behavior of Bayesian Dirichlet-Laplace shrinkage (Bhattacharya et al., 2015). It is worth to note that if certain minimal signal strength condition holds as well, i.e., $\min_{\{\vartheta_i^* \neq 0\}} |\vartheta_i^*|$ is bounded away from zero, with additional assumptions, one can derive an even stronger posterior selection consistency result that $\pi(G(\theta, \sigma) = G^*|y) \to^p 1$, following the proof of Theorem 2.3 in Song and Liang (2017). Readers of interests can easily derive such posterior selection consistency by themselves.

It is not difficult to verify the condition (2.5) for the proposed t fusion shrinkage prior (2.4).

Corollary 2.1. *When* f_θ *is the cdf of normal distribution* $N(0, \lambda_1)$ *and* f_ϑ *is the cdf of a* t *distribution with scale parameter* s*, then the first three inequalities of* Eq. 2.5 *hold when*

$$\log(\max_i |\vartheta_i^*|/\sigma^*) = O(\log n),$$
$$\theta_1^{*2}/(\lambda_1 \sigma^{*2}) + \log(\lambda_1) = O(\log n),$$

and $s = n^{-c}$ *for some sufficiently large* c*.*

The above results hold not only for the scaled t prior, but also for any other choice of f_ϑ, as long as f_ϑ has a polynomially decaying tail. Note that the above results can not be generalized to light tailed distributions such as Laplace fusion prior, since it will lead to an unrealistic sufficient condition that $\max_i |\vartheta_i^*|/\sigma^* = o(1)$.

2.3. Bayesian Posterior Inference In this section, we illustrate the posterior behavior of Bayesian t fusion with a toy example, and compare it to Bayesian Laplace fusion. We will also discuss other issues related to hyper-parameter choice.

A simulation data was generated with $n = 100$ and $\sigma^* = 0.5$. The data and underlying true parameter value are plotted in Fig. 2. Three estimation

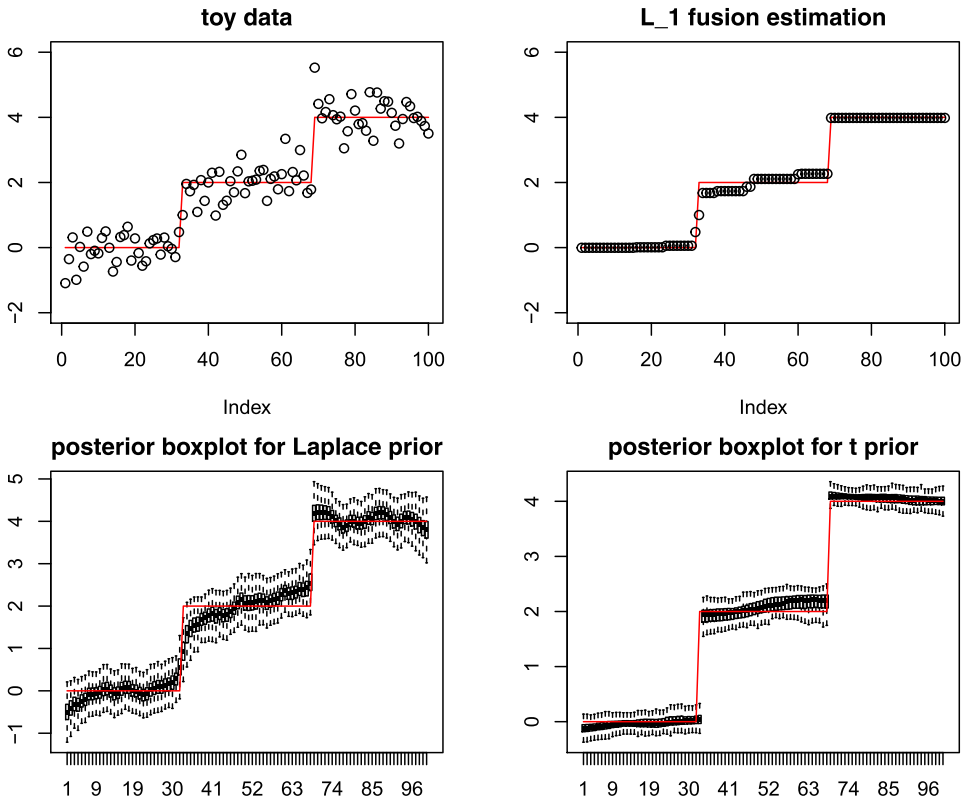

Figure 2: Upper Left: Simulated toy data; Upper Right: Frequentist fusion estimation (2.1); Lower Left: Marginal posterior boxplots for each θ_i under Laplace fusion prior; Lower Right: Marginal posterior boxplots under t fusion prior. The red curve denotes the true θ_i^*'s which contain three blocks

procedures are considered: 1) L_1 fusion estimation (2.1) where the tuning parameter is selected by cross validation; 2) Bayesian Laplace prior (2.2) with $\lambda = \sqrt{2\log n}$, $a_\sigma = 0.5$, $b_\sigma = 0.5$ and λ_1=5; 3) Bayesian t prior (2.4) with $a_t = 2$, $b_t = 0.005$, $a_\sigma = 0.5$, $b_\sigma = 0.5$ and $\lambda_1 = 5$. Note the reason we choose $\lambda = \sqrt{2\log n}$ for Bayesian Laplace is that the theoretical choice for the tuning parameter in frequentist L_1 fused estimation is of order $\sigma^*\sqrt{\log n}$ (Rinaldo et al., 2009).

The posterior samples are obtained by Gibbs sampler for 2000 iterations. The frequentist estimator and boxplots of Bayesian marginal posterior distributions are displayed in Fig. 2. Note that for the sake of readability of the figure, we don't draw the outliers in these boxplots.

The comparison shows that the L_1 fusion penalty leads to a strictly sparse $\widehat{\vartheta}$ estimation and the estimated $\widehat{\theta}$ does have a blocking structure, but there is a mild over-partition issue. Due to the nature of Bayesian shrinkage prior, the two Bayesian approaches only produce continuous posterior samples of ϑ_i. Compared with the Laplace prior, the advantages of t prior are quite obvious. Its posterior concentration is better, i.e., shorter boxes and whiskers for the boxplots, and posterior mean is much closer to the true red curve. Although its posterior estimation is not exactly blocky, one can visually identify the three blocks. In contrast, the Laplace prior generates larger posterior variance, and its posterior center smoothly fluctuates around the true step function. As discussed at the end of Section 2.1, Laplace fusion prior itself merely introduces shrinkage effect for the successive differences, and hence the data fluctuation carries over to the posterior distribution of θ. In this case, it is clear that it has not recovered the underlying block structure.

Note that for both (2.2) and (2.4) prior specifications, their posterior distributions highly depend on the choices of hyperparameters, especially the scale parameters of the Laplace (parameter $1/\lambda$) and t prior (parameter s) distribution. To understand the influence of the scale parameter, we increase and decrease the scale parameter by a factor of $\sqrt{10}$, for both Laplace and t priors. The corresponding posteriors are displayed in Fig. 3.

From Fig. 3 and its comparison with Fig. 2, we see that, for t prior specification, smaller scale hyperparameter leads to smaller posterior variation, and vise versa. Choosing an overly large scale parameter weakens the shrinkage effect, thus it fails to shrink the θ_i's that belong to the same block towards same value, and the corresponding posterior estimation is not flat within blocks. On the other hand, choosing an overly small scale hyperparameter, although yields very strong shrinking and grouping effect, may potentially over-partition the data. In conclusion, the scale parameter of t fusion prior controls the aggressiveness of posterior block partition. For Laplace prior, both increasing or decreasing scale parameter can not remedy its poor behavior demonstrated in Fig. 2. The choice of scale parameter only affects the smoothness of posterior of θ. Smaller scale parameter leads to smoother nonparametric estimation which is similar to smooth spline regression; larger scale parameter leads to a rather rugged estimation. This phenomenon can be partially explained by Fig. 1. As the scale of the Laplace prior decreases (i.e., λ increases), the conditional prior of θ_i given $(\theta_{i-1}, \theta_{i+1})$ acts more and more like a uniform distribution over $[\theta_{i-1}, \theta_{i+1}]$. Hence, conditional on $(\theta_{i-1}, \theta_{i+1})$, the posterior of "local" total variation, $|\theta_{i+1} - \theta_i| + |\theta_i - \theta_{i-1}|$, always decreases to the minimum possible value (i.e., $|\theta_{i+1} - \theta_{i-1}|$) when the scale parameter decreases to zero, regardless of the observation value y_i.

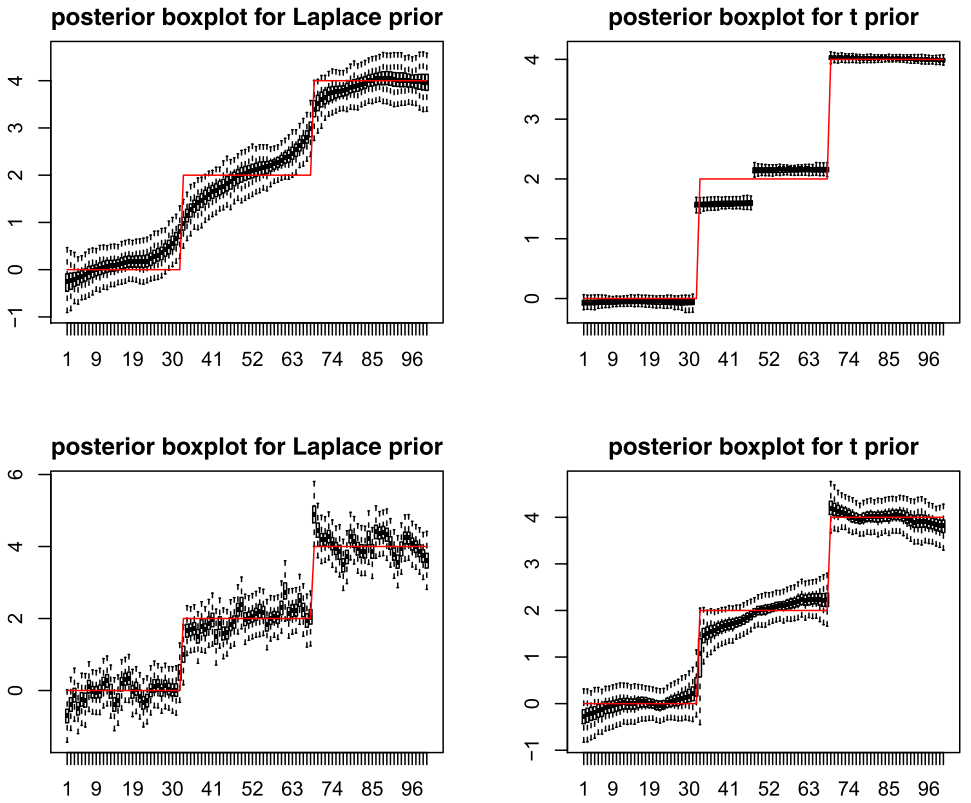

Figure 3: Bayesian posterior with different choice of scale parameter. Upper Row: The scale parameter decreases by a factor of $\sqrt{10}$. Lower Row: The scale parameter increases by a factor of $\sqrt{10}$

As showed in the previous toy example, the hyperparameter value plays an important role for the performance of Bayesian posterior inferences. The theoretical suggestion, i.e., the first inequality of Eq. 2.5, gives a very small scale parameter, which in practice leads to severe over-partition issue under moderate n. Unlike frequentist high dimensional penalization estimation whose tuning parameter is usually determined by cross validation or selection criterion such as EBIC (Chen and Chen 2008, 2012), choosing an appropriate value for the hyperparameter posts many difficulty for Bayesian statisticians. Conventional choices include imposing a hyper-prior on the hyperparameter (van der Pas et al., 2017; Castillo and van der Vaart, 2012) or empirical Bayes (Robbins, 1985). For high dimensional sparse GLMs, Liang et al. (2013) suggests choosing a hyperparameter such that the posterior mean and posterior mode are close. Several Bayesian works (Shimamura et al., 2018;

Betancourt et al., 2017) consider only using the sparse MAP as the Bayesian estimator, and thus abandon the whole posterior distributional information. This strategy somehow reduces the Bayesian computation to a frequentist optimization problem, and then the hyperparameter can be determined by EBIC criterion. It is beyond the scope of this work to theoretically study how to select an appropriate hypereparameter, i.e., the scale parameter of the t prior. An empirical suggestion based on authors' experience is to choose the scale parameter of Eq. 2.4 such that

$$P(|t_{df}(s)| \geq \sqrt{\log n/n}) \approx 1/n. \tag{2.6}$$

Although it doesn't quite satisfy the conditions in Corollary 2.1 which suggests $s = n^{-c}$ for some large c, but in practice, it indeed yields a reasonable and stable Bayesian performance. Note that our prior choice $a_t = 2$, $b_t = 0.005$ in the above toy example approximately satisfies (2.6).

3 Bayesian t Shrinkage for Bayesian Clustering

In this section, we would like to extend the applications of Bayesian t fusion shrinkage to Bayesian clustering problem, where the parameter θ in model (1.1) is assumed to follow an unknown clustering structure. In other words, the observations y_i's are not organized in a sensible order, and we no longer assume that the true cluster consists of only consecutive indexes.

A full Bayesian clustering analysis, or subgroup identification, usually imposes a prior on the clustering structure, including the number of clusters and how to partition observations into these clusters. For example, one could consider that the cluster structure arises from a tree splitting process, and impose certain prior on the tree nodes (Berger et al., 2014; Heller and Ghahramani, 2005). The resultant posterior sampling hence usually requires a reversible-jump Metropolis-Hastings move to travel across different clustering models in the sampling space, which can be quite inefficient. Another common approach is to model a mixture distribution via nonparametric priors such as Dirichlet process or Chinese restaurant process (Neal, 2000). These mentioned approaches enforce explicit *clustery* posterior samples by directly applying discrete prior on subgroup structure. In this section, we will show how to use Bayesian t shrinkage to induce implicit posterior clustering structure.

To implement shrinkage prior, we need to formulate the problem in a proper statistical modeling framework such that its parameter possesses certain sparsity feature. Conceptually, imposing shrinkage priors on $\theta_i - \theta_{i-1}$ shall not work for this clustering problem since $\{\theta_i - \theta_{i-1}\}_{i=2}^n$ is no longer a sparse vector. However, if some meaningful permutation r of θ_i is known

such that $\{\theta_{r(i)} - \theta_{r(i-1)}\}_{i=2}^n$ is indeed a sparse vector, for example, $\theta_{r(i)}$ is ascending or descending, then the problem will reduce to Bayesian shrinkage fusion studied in the previous section.

However, in practice, such $r(\cdot)$ is generally not available. Thus one simple solution would be to substitute $r(\cdot)$ with some estimator $\widehat{r}(\cdot)$. For model (1.1), a trivial estimator is the rank statistic of the responses y_i, i.e., $y_{\widehat{r}(i)} = y_{(i)}$ where $y_{(i)}$ denotes the order statistics of y_i's. Thus this leads to hybrid Bayesian modeling, depending on a frequentist pilot estimator $\widehat{r}$:

$$\begin{aligned} \sigma^2 &\sim \text{Inverse-Gamma}(a_\sigma, b_\sigma), \quad \theta_{\widehat{r}(1)}|\sigma^2 \sim N(0, \sigma^2\lambda_1), \\ (\theta_{\widehat{r}(i)} - \theta_{\widehat{r}(i-1)})|\sigma^2, \lambda_i &\sim N(0, \sigma^2\lambda_i), \quad \lambda_i \sim \text{Inverse-Gamma}(a_t, b_t) \quad \text{for all } i = 2, \ldots, n. \end{aligned} \tag{3.1}$$

The corresponding posterior Gibbs sampler follows:

$$\begin{aligned} \lambda_i &\sim \text{Inverse-Gamma}\left(a_t + 1/2, b_t + \frac{(\theta_{\widehat{r}(i)} - \theta_{\widehat{r}(i-1)})^2}{2\sigma^2}\right) \text{ for all } i = 2, \ldots, n. \\ \sigma^2 &\sim \text{Inverse-Gamma}\left(a_\sigma + n, b_\sigma + \frac{\|y-\theta\|^2}{2} + \frac{\theta_{\widehat{r}(1)}^2}{2\lambda_1} + \sum_{i=2}^n \frac{(\theta_{\widehat{r}(i)} - \theta_{\widehat{r}(i-1)})^2}{2\lambda_i}\right), \\ \theta_{\widehat{r}(i)} &\sim N(\mu_i, \nu_i), \quad \text{for } i = 1, \ldots, n \end{aligned} \tag{3.2}$$

where $\nu_i^{-1} = 1/\sigma^2 + 1/\lambda_{i+1}\sigma^2 + 1/\lambda_i\sigma^2$ and $\mu_i = \nu_i(y_{\widehat{r}(i)}/\sigma^2 + \theta_{\widehat{r}(i+1)}/\lambda_{i+1}\sigma^2 + \theta_{\widehat{r}(i-1)}/\lambda_i\sigma^2)$.

The idea of taking advantage of a pilot estimator $\widehat{\theta}$ and its rank statistic $\widehat{r}(i)$ has also been implemented in the two-stage frequentist approach (Ke et al., 2015a).

It is worth to mention that posterior consistency results described in Theorems 2.1 and 2.2 for Bayesian shrinkage fusion don't apply to the above Bayesian modeling (3.1) even when $\theta_{\widehat{r}(i)}$ is in ascending order, because $\varepsilon_{\widehat{r}(i)}$'s are no longer iid error observations. Such a data dependent prior (3.1) can also be interpreted as a misspecified Bayesian modeling (Kleijn et al., 2006a), where the data is the order statistics $y_{(i)}$ from a mixture distribution, and we model $y_{(i)}$'s as independent Gaussian variables whose mean function $E(y_{(i)})$ is a step function.

Although the prior modeling (3.1) is quite natural and intuitive, there are several issues. First, compared with the iid observations, it is more difficult to to identify the underlying clustering structure of sorted observations. For example, it is much more difficult to visually identify the 3-cluster structure in the toy data showed in Fig. 4 than the toy data showed in Fig. 2. Secondly,

Eq. 3.1 fails to take account of uncertainty of the estimator $\widehat{r}$, and the strict monotonicity of $y_{\widehat{r}(i)}$ will always carry over to the posterior of $\theta_{\widehat{r}(i)}$. Thirdly, over-clustering will occur. To understand this, let us consider that all data y_i's have the same mean, i.e., there is only one cluster. But the mean function of the sorted data, $E(y_{(i)})$, is actually a strictly increasing function. From the perspective of Bayesian misspecification, our posterior of $\theta_{\widehat{r}(i)}$'s should contract towards the best step function, in the sense of minimizing the Kullback–Leibler divergence (Kleijn et al., 2006a), rather than towards the constant function $E(y_i)$. This causes the over-clustering, and furthermore, the posterior consistency of t shrinkage prior established in Corollary 2.1 doesn't hold anymore.

Theorem 3.1. *Assume that all* $y_i \sim N(0,1)$ *with known variance* $\sigma^* = 1$. *If the prior* (3.1) *is used (except that there is no necessity to impose a prior on* σ^2*) and the scale parameter satisfies* $-\log(b_t) \asymp \log n$, *then in high probability*

$$\pi(\|\theta\| \le \sqrt{M \log n}|y) < 1/2,$$

i.e., the rate-$\sqrt{\log n}$ *posterior consistency fails.*

The negative result in Theorem 3.1 motivates us to propose an adaptive pseudo Bayesian shrinkage approach. Instead of using a fixed estimator $\widehat{r}$, we update the "working" order r of θ over iterations. To be specific, we adopt the following pseudo "Gibbs" sampling algorithm: In each iteration,

$$\begin{aligned}
&1.\ \text{update } \lambda_i \sim \text{Inverse-Gamma}\left(a_t + 1/2, b_t + \frac{(\theta_{r(i)} - \theta_{r(i-1)})^2}{2\sigma^2}\right) \text{ for all } i = 2, \ldots, n.\\
&2.\ \text{update } \sigma^2 \sim \text{Inverse-Gamma}\left(a_\sigma + n, b_\sigma + \frac{\|y-\theta\|^2}{2} + \frac{\theta^2_{r(1)}}{2\lambda_1} + \sum_{i=2}^{n} \frac{(\theta_{r(i)} - \theta_{r(i-1)})^2}{2\lambda_i}\right),\\
&3.\ \text{update } \theta_{r(i)} \sim N(\mu_i, \nu_i), \quad \text{for } i = 1, \ldots, n,\\
&4.\ \text{update } r(\cdot) = \text{the rank statistic of the current sample } \theta.
\end{aligned} \tag{3.3}$$

where in step 3, $\nu_i^{-1} = 1/\sigma^2 + 1/\lambda_{i+1}\sigma^2 + 1/\lambda_i\sigma^2$ and $\mu_i = \nu_i(y_{r(i)}/\sigma^2 + \theta_{r(i+1)}/\lambda_{i+1}\sigma^2 + \theta_{r(i-1)}/\lambda_i\sigma^2)$. Comparing to Eq. 3.2, the essential difference is that we update $r(\cdot)$ to the rank statistic of the newly obtained θ in each iteration.

There are a couple of rationales behind the algorithm (3.3). First, the update of r potentially allows algorithm (3.3) to incorporate some uncertainty of rank statistic into the posterior sampling. Heuristically, if the current $r(\cdot)$ is close to the true ranking, Eq. 3.3 will shuffle the ordering of θ within each cluster, hence $y_{r(i)}$ acts like independent samples yielded by random

permutation within cluster, rather than order statistics. Secondly, the sampling algorithm (3.3) can be somehow connected with a full Bayesian modeling:

$$\begin{aligned} r &\sim \text{Uniform distribution over all possible permutations,} \\ \sigma^2 &\sim \text{Inverse-Gamma}(a_\sigma, b_\sigma), \quad \theta_{r(1)}|\sigma^2 \sim N(0, \sigma^2\lambda_1), \\ (\theta_{r(i)} - \theta_{r(i-1)})|\sigma^2, \lambda_i &\sim N(0, \sigma^2\lambda_i), \quad \lambda_i \sim \text{Inverse-Gamma}(a_t, b_t) \quad \text{for all } i = 2, \ldots, n. \end{aligned} \tag{3.4}$$

Under (3.4), the conditional posterior of $r(\cdot)$ follows $\pi(r|\theta, \lambda_1, \ldots, \lambda_n, y) \propto \exp\{-\sum_{i=2}^{n}(\theta_{r(i)} - \theta_{r(i-1)})^2/\lambda_i^2\sigma^2\}$, from which sampling is expensive. However, when most of λ_i^2s are tiny, the distribution of $\pi(r|\theta, \lambda_i's, y)$ will be highly concentrated around its MAP which is approximately the rank statistic of θ. Therefore, updating r to the rank statistic of current sample θ, as in Eq. 3.3, can be viewed as a convenient sampling of $\pi(r|\theta, \lambda_i's, y)$ under the uniform hyperprior of r. And such a hyperprior does benefit the posterior asymptotic performance, at least it remedies the posterior inconsistency described in Theorem 3.1 when there is only one underlying cluster.

Theorem 3.2. *Assume that all* $y_i \sim N(\theta_0, \sigma^2)$, *i.e.* θ^* *is a vector of* θ_0*'s. If prior* (3.4) *is used with* $b_t = n^{-c}$ *for some sufficiently large* c, *and* $\theta_0^2/(\lambda_1\sigma^{*2}) + \log(\lambda_1) = O(\log n)$ *then there exist constant* M *and* $\epsilon_n \asymp \sqrt{\log n/n}$, *such that*

$$\pi(\|\theta - \theta^*\| \geq M\sigma^*\sqrt{n}\epsilon_n|y) \to 0,$$

where the convergence holds in probability and L_1.

Further investigations will be pursued to assess the posterior contraction induced by Eq. 3.4 when the true parameter θ^* has multiple-cluster structure. The minimal cluster difference $\min_{\{(i,j):\theta_i \neq \theta_j\}} |\theta_i - \theta_j|$ shall play a crucial role for the Bayesian asymptotics. If the minimal cluster difference is bounded by constant, then even the best Bayes classifier makes as many as $O(n)$ misclassifications, which leads to $\sqrt{n}$-rate L_2 error. And we conjecture that when the minimal cluster difference is large than $\sqrt{M \log n}$ for some constant M, the posterior distribution induced by Eq. 3.4 leads to satisfactory Bayesian consistency.

It is worth to mention that the r-update step in algorithm (3.3) may critically change the stochastic stability of the algorithm and the ergodicity of Markov chains generated by Eq. 3.3 is still unclear. Therefore, in our application we use it with caution. In all our simulations and toy examples, we initialize with $r = \widehat{r}$, and the r-update step (i.e. step 4 in Eq. 3.3) is only implemented every other 20 iterations after certain burn-in period.

In frequentist literature, another popular approach to cluster θ_i's is to impose a pairwise difference penalty as $\sum_{i\neq j} p(|\theta_i-\theta_j|)$ for some penalty $p(\cdot)$ (Ma and Huang, 2017; Shen and Huang, 2012; Ke et al., 2015b). Although it is tempting to develop a Bayesian counterpart, i.e., using a prior of form $\pi(\theta) \propto \prod_{i\neq j} \pi(\theta_i - \theta_j)$ with sparsity induced π, there are several problems. First, the pairwise differences $\{\theta_i - \theta_j\}_{i\neq j}$ must satisfy triangle inequality, hence the prior specification $\pi(\theta) \propto \prod_{i\neq j} \pi(\theta_i - \theta_j)$ is actually not an independent prior over all pairwise difference $\{\theta_i - \theta_j\}_{i\neq j}$, and it will be difficult to characterize the effect of such dependency in the prior distribution. Second, even under clustering structure, $\{\theta_i - \theta_j\}_{i\neq j}$ are actually not sparse. For example, if there are two balanced groups among θ_i's, then half of the pairwise differences are nonzero. Imposing a sparse prior on a non-sparse system will lead to severe overshrinkage problem, and our simulation shows that such prior shrinks all θ_i towards $\bar{y}$. Third, such pairwise difference prior will also heavily increase the computational burden of posterior sampling.

To illustrate and compare the performance of Eqs. 3.2 and 3.3, a simple toy example is conducted. We simulate a data set with $n = 100$ data points that belong to 3 underlying subgroups. For both (3.2) and (3.3), we choose $a_t = 2$, $b_t = 0.005$, $a_\sigma = b_\sigma = 0.5$ and $\lambda_1 = 5$. In addition, we compare them with frequentist L_1 fusion

$$\widehat{\theta} = \arg\min \|y-\theta\|^2/2 + \lambda \sum_{i=2}^{n} |\theta_{\widehat{r}(i)} - \theta_{\widehat{r}(i-1)}|, \tag{3.5}$$

and Bayesian Dirichlet process modeling

$$\begin{aligned} &y_i|\theta_i, \sigma^2 \sim N(\theta_i, \sigma^2), \quad \theta_i \sim G, \\ &G \sim \text{Dirichlet-Process}(N(0,\lambda), a) \text{ and } \sigma^2 \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma) \end{aligned} \tag{3.6}$$

with $\lambda = 5$, $a = 0.1$ and $a_\sigma = b_\sigma = 0.5$. The posterior sampling algorithms of Dirichlet process modeling are discussed in (Neal, 2000). The frequentist estimator and posterior boxplots of θ_i's under different priors are plotted in Fig. 4. This figure clearly shows that frequentist L_1 fusion estimator fails to yield clustering structure for monotone sorted $y_{\widehat{r}(i)}$'s, especially for the middle portion of the data, one always has $\widehat{\theta}_{\widehat{r}(i)} = y_{(i)}$. Dirichlet process induces a quite reasonable posterior clustering, however its posterior concentration is not good, in the sense that the posterior variance is quite large. In the opposite, the t modeling (3.1) has a strong prior concentration, which contributes to a better posterior variation, and encourages posterior clustering. However, consistent to our previous arguments, a severe over-clustering

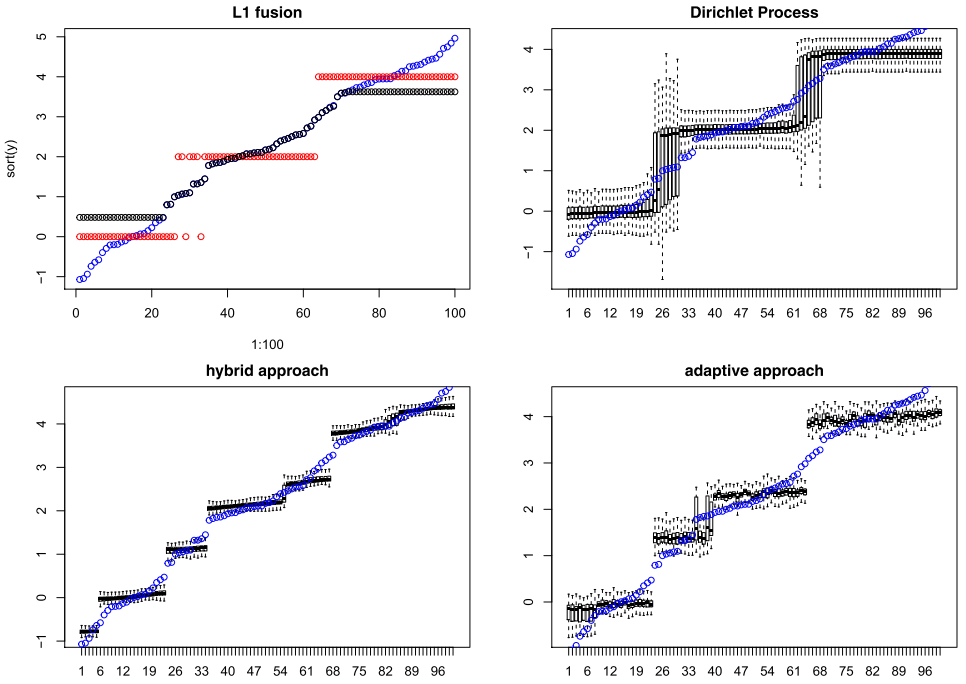

Figure 4: Toy example comparison among difference approaches. Upper Left: Frequentist fusion estimation (3.5); Upper Right: Marginal posterior boxplots for each θ_i under Dirichlet process prior; Lower left: Marginal posterior boxplots for each θ_i under t prior (3.2); Lower right: Marginal posterior boxplots under t prior (3.3). The blue points denote the sorted observed data $y_{(i)}$. The red points denote the true θ^* corresponding to the data ascending order

occurs, and the posterior identifies more than 6 clusters. At last, the posterior obtained by algorithm (3.3) combines both advantages of DP prior and Eq. 3.1. Comparing with DP prior, it has much smaller posterior variance; comparing with prior (3.1), the over-clustering issue is greatly alleviated. More discussions can be found in the simulation section.

4 Simulation and Data Anlaysis

4.1. Bayesian Fusion Simulations In our first simulation studies, we consider model (1.1) with $\sigma^* = 0.5$ and true parameter $\theta^* \in \mathbb{R}^{100}$ having three consecutive blocks. These three blocks are $\mathcal{B}_1^* = \{1, \ldots, b_1\}$, $\mathcal{B}_2^* = \{b_1 + 1, \ldots, b_1 + b_2\}$, $\mathcal{B}_3^* = \{b_1 + b_2 + 1, \ldots, b_1 + b_2 + b_3\}$ where $(b_1, b_2, b_3) \sim$ multinomial$(100, (1/3, 1/3, 1/3))$, and $\theta_i^* = 0$, 2 or 4 within

each block respectively. We compare the performance among Bayesian t fusion (2.4), Bayesian Laplace fusion (2.2) and frequentist L_1 fusion (2.1), based on 100 replications. The choices of the hyperparameters are same to the toy example discussed in Section 2.3.

To compare the accuracy of parameter estimations, we calculate the L_2 estimation error, $\|\widehat{\theta} - \theta^*\|_2$ and $\|\widehat{\theta} - \theta^*\|_1$, for frequentist and Bayes estimator (posterior mean). To assess the performance of block structure recovery, we consider several comparison criteria. We define the "adjacency" matrix as $\Sigma = (\omega_{ij}) = (1\{\theta_i = \theta_j\})$, and corresponding estimation error of Σ: $R = \|\Sigma^* - \widehat{\Sigma}\|_1 = \sum_{i,j} |\widehat{\omega}_{ij} - \omega^*_{ij}|$. For L_1 fusion estimator, $\widehat{\Sigma}$ is trivially estimated by $\widehat{\omega}_{ij} = 1\{\widehat{\theta}^{\mathrm{F}}_i = \widehat{\theta}^{\mathrm{F}}_j\}$. For Bayesian shrinkage approaches, since their posterior samples are continuous, it is necessary to "sparsify" the continuous posterior in order to retrieve a sparse structure estimation for Ω. In literature, this is usually done by 1) hard thresholding (Li and Pati, 2017; Tang et al., 2016; Carvalho et al., 2010; Ishwaran and Rao, 2005), or 2) decoupling shrinkage and selection methods (Hahn and Carvalho, 2015; Xu et al., 2017). All these mentioned approaches depend solely on the posterior scaling, and do not take into account of the degree of prior concentration. Therefore, following the suggestion made in Song and Liang (2017), we estimate $\widehat{\omega}_{ij} = 1\{|\widehat{\theta}_i - \widehat{\theta}_j|/\widehat{\sigma} \leq \pi_{1/2n}\}$, where $\widehat{\theta}_i$ and $\widehat{\sigma}$ are the Bayes estimator, $\pi_{1/2n}$ denotes the $(1 - 1/2n)$ quantile of the prior distribution imposed on successive difference ϑ_i/σ. This choice tries to increases the robustness of Bayesian inference: if an overly small or large scale parameter s is used, the estimation for $\omega_{i,j}$ will adapt accordingly. For t shrinkage (2.4), $\pi_{1/2n} = st_{1/2n}$ where s is the scale parameter, $t_{1/2n}$ is the $1 - 1/2n$ quantile of t distribution with $\mathrm{df} = 2a_t$; for Laplace shrinkage (2.2), $\pi_{1/2n} = \log n/\lambda$. Since different priors have different corresponding $\pi_{1/2n}$ values, and comparison solely based on the value of R may not be completely fair. Hence, we also consider the following two statistics: $W = \mathrm{Average}_{\{w^*_{ij} \neq 0\}} |\widehat{\theta}_i - \widehat{\theta}_j|$ denotes within-block average variation, and $B = \min_{\{w^*_{ij}=0\}} |\widehat{\theta}_i - \widehat{\theta}_j|$ denotes the between-block separation. A larger B and smaller W indicate a better block identification performance. The results are summarized in Table 1.

The simulation results show that the Bayesian t fusion yields the most accurate estimation in terms of L_2 error. This Bayes estimator also induces the best clustering result, as it has the largest between-group separation and smallest within-group variation. In comparison, the frequentist fusion estimator has a worse accuracy. It is well known that L_1 penalty introduces estimation bias (Fan and Li, 2001), but unlike the LASSO penalty which introduces a bias of λ, fused L_1 penalty introduces only a bias as small as

Table 1: Comparison among Bayesian t fusion, Bayesian Laplace fusion, and L_1 fusion

Methods	Bayesian t fusion	Bayesian Laplace fusion	L_1 fusion
L_2 error of θ	1.3243	2.2323	1.5916
Standard error	0.0602	0.0324	0.0421
W (the smaller the better)	0.05584	0.2182	0.2128
Standard error	0.0028	0.0037	0.0077
B (the larger the better)	1.4243	0.6792	1.1302
Standard error	0.0684	0.0211	0.0398
R (the smaller the better)	387.32	85.6	1360.9
Standard error	32.0628	6.2375	53.2685

The report is based on average results from 100 replications. Refer to Section 4.1 for the definition of R, W and B

λ/n_i (Rinaldo et al., 2009) where n_i is the number of elements in each block. Hence, its estimation performance is still acceptable. The L_1 fusion also has a much larger R statistics. This is consistent to observation from our toy example, that L_1 fusion has a mild over-partition issue. Bayesian Laplace fusion, on the other hand, has a much worse estimation behavior due to its insufficient shrinkage effect. As discussed in Section 2.3, the Laplace fusion tends to yield smooth changing $\widehat{\theta}_i$'s, thus it has a much smaller between-block separation. Note that Laplace shrinkage does obtain the smallest R statistic, but this doesn't imply that it has a good performance on structure recovery. It has a small R statistics because the Laplace fusion prior doesn't have strong prior concentration and its corresponding $\pi_{1/2n}$ is much larger than t fusion.

4.2. Bayesian Clustering Simulations In our second simulation studies, we consider model (1.1) with $\sigma^* = 0.5$ and true parameter $\theta^* \in \mathbb{R}^{100}$ having three unknown clusters. With equal probability, $\theta_i^* = 0$, 2, or 4. We aim to compare different approaches including Bayesian t modeling (3.2) and (3.3), Dirichlet process prior (3.6) and the frequentist bCARDS estimation (3.5) using L_1 fusion (Ke et al., 2015a). Besides the comparison of L_2 error of the Bayesian/frequentist estimator, we also compare the posterior mean of squared L_2 error, $E_{\pi(\theta|y)}\|\theta - \theta^*\|^2$, among the Bayesian approaches. Note that the L_2 error of Bayes estimator only tells how good is the posterior center, while the posterior mean of squared L_2 error tells how good is distributional posterior contraction.

Note that the simulation setting seems similar to our previous study in Section 4.1, but as mentioned in Section 3, clustering problem is much more difficult than fusion problem. To see that, assume there are three balanced clusters for θ, let z_1 be the largest observation in the 0's cluster and z_2 be the smallest observation in the 2's cluster. By the well known result (Royston, 1982), $E(z_1) \approx 1.04 > 0.96 \approx E(z_2)$. Therefore, misclassification will always happen for those extreme observations, and $B = \min_{\{w_{ij}^*=0\}} |\widehat{\theta}_i - \widehat{\theta}_j|$ is always around 0 for all methods, therefore, the comparison of B statistic values are meaningless. Because of that, to assess how well these methods identify and separate unknown clusters, we re-define the B statistic as

$$\tilde{B} = \text{the bottom } 10\% \text{ quantile of } \{|\widehat{\theta}_i - \widehat{\theta}_j|\}_{\{w_{ij}^*=0\}}.$$

The simulation results are summarized in Table 2.

The comparison shows that DP prior yields the best point estimator in terms of estimation error, and the adaptive t shrinkage method (3.3) gives a slightly worse, but comparable point estimator. t shrinkage approaches do yield better posterior contraction than DP prior (i.e., smaller posterior mean of squared L_2 error), which is consistent to the insight obtained from the toy example in Section 3.

The three Bayesian approaches have approximately the same performance for between-cluster separation ($\tilde{B}$ statistic), while the DP prior has a smaller within-cluster variation (W statistic). The frequentist estimator (3.5) has the worst performance in almost every aspect. In summary, DP prior does yield the best Bayes point estimation, but the adaptive t shrinkage (3.3) has the best Bayesian contraction, and its posterior mean, as a point estimator, has reasonable performance for estimation and clustering structure recovery. Furthermore, by comparing the results between t shrinkage method (3.3) and (3.2), we conclude that the r-update step implemented in algorithm (3.3) does improve the performance of t shrinkage in every aspect for this clustering problem.

It is worth to remind the readers that although the adaptive t fusion approach (3.3) updates the working order r over iterations, it still doesn't take into account full uncertainty of r, since the r-update step ignores all other possible choices of data ordering. As pointed out by one of the referees, this greatly reduces the posterior uncertainty, and leads to underestimated posterior variance. This doesn't affect the Bayesian estimation accuracy too much, as seen from the simulations, but will affect the validity of Bayesian high-order inferences such as Bayesian credible intervals. Our simulation shows that under 90% nominal level, the credible intervals induced by the DP prior achieves an average of 90.71% coverage, but the adaptive t fusion

Table 2: Comparison among Bayesian t shrinkage, Bayesian DP prior, and L_1 fusion

Methods	t shrinkage (3.3)	t shrinkage (3.2)	DP prior (3.6)	L_1 fusion (3.5)
posterior mean of L_2^2 error	20.559	20.5660	40.7029	—
Standard error	0.6885	0.4589	3.1209	—
L_2 error of θ	4.1873	4.3925	3.9584	5.0039
Standard error	0.0872	0.0531	0.1621	0.0739
W (the smaller the better)	0.3517	0.4428	0.2183	0.4551
Standard error	0.0093	0.0063	0.0061	0.0062
$\tilde{B}$ (the larger the better)	1.4652	1.3392	1.3885	0.9855
Standard error	0.0218	0.0147	0.0419	0.0146
R (the smaller the better)	1752.24	2294.76	—	1973.7
Standard error	31.5641	14.6102	—	11.9937

The report is based on average results from 100 replications. Refer to Section 4.2 for the definition of R, W and $\tilde{B}$

and hybrid t fusion approaches only achieve 55% and 23% coverage probability respectively. Therefore, further theoretical investigations on posterior properties, such as Bernstein-von Mises phenomenon, are necessary.

4.3. Real Data Set Analysis In this section, we consider a real comparative genomic hybridization (CGH) dataset (Tibshirani and Wang, 2007). The dataset is available from the R package `cghFLasso` and it contains $n = 990$ observations. We are interested in a fusion estimation for the mean trend in order to detect the "hot-spot" region, i.e., the corresponding genes have extra DNA copies. We apply Bayesian t shrinkage fusion (2.4),

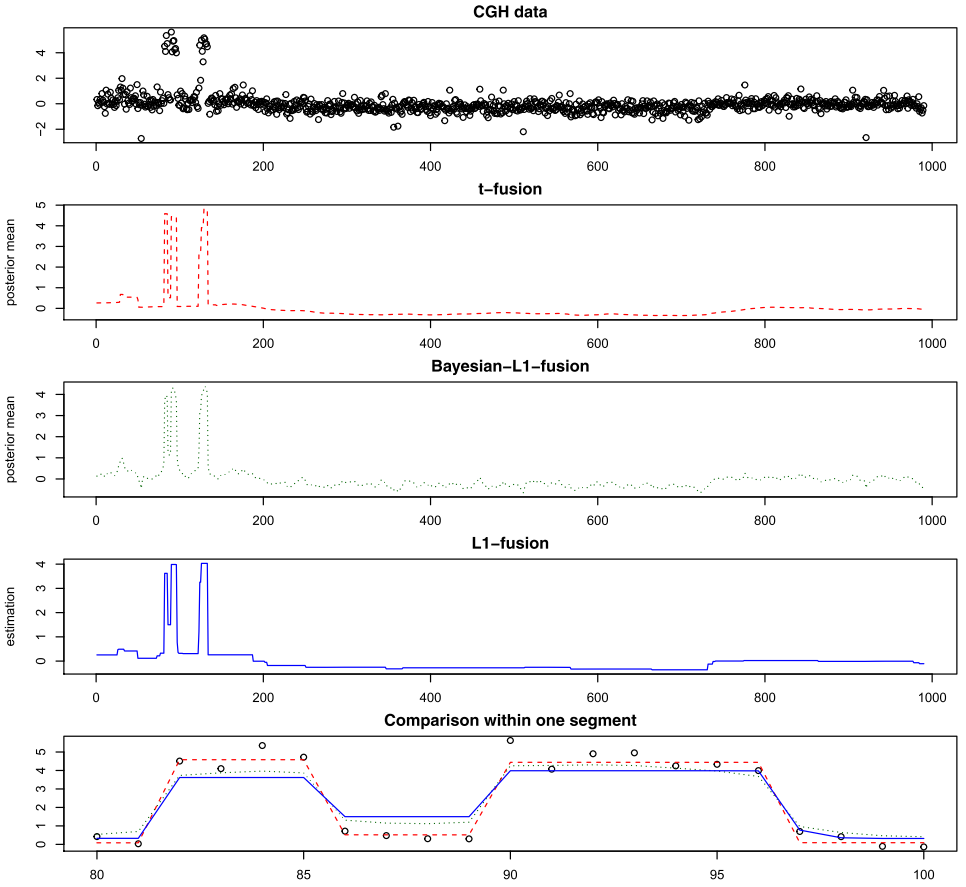

Figure 5: Real data application: 1) CGH data; 2) posterior mean of t shrinkage fusion; 3) posterior mean of Bayesian Laplace fusion; 4) frequentist L_1 fusion estimator; 5) comparison of the two approaches within a segment of [80,120]

Bayesian Laplace fusion, and frequentist L_1 fusion (2.1) to this real data set. For the Bayesian t fusion approach, the hyperparameter is chosen as $a_t = 2$ and $b_t = 0.0007$, following the suggestion of Eq. 2.6. For the Bayesian Laplace prior, its $\lambda = \sqrt{2\log(990)}$. All these point estimations are plotted in Fig. 5. The Bayesian Laplace fusion method yields a rather rugged mean trend estimation, and in contrast, the Bayesian t fusion and frequentist L_1 fusion yield a much flatter and smoother mean trend estimation. To compare Bayesian t fusion and frequentist L_1 fusion approaches and demonstrate the advantage of t shrinkage, we display the estimation results within the segment [80,120]. It is obvious that all approaches induce almost the same blocking structure. But the difference is that, for t shrinkage prior, the $\widehat{\theta}_i$ within each block is very close to the sample mean of the block; but for L_1 fusion and Bayesian Laplace fusion, there is a large bias between $\widehat{\theta}_i$ and the sample mean of the block.

5 Final Remarks

In this work, we study the Bayesian inference for a vector parameter θ which has an unknown blocking structure, using Bayesian shrinkage prior. For the ease of representation, this work focuses on the application of t fusion prior, but our theorems actually holds for any polynomially decaying prior distribution. We demonstrate that a simple t fusion prior leads to satisfactory posterior contraction, and is powerful to recover the blocking structure. We recommend not to use Laplace fusion prior since it can only induce a smoothly varying posterior estimation. Although this work mainly focuses on the normal sequence models, but the presented t fusion modeling and the posterior asymptotic results can be easily extended to more complicated regression models such as $y = X\beta + Z\theta + \epsilon$ for some blocky parameter θ.

We also extend the use of shrinkage prior to a more general clustering problem. To the best of authors' knowledge, this is the first attempt in literature to use Bayesian shrinkage to recover unknown clustering structure. The basic idea is to find a pilot order $\widehat{r}$, and then fuse all pairs of neighbors (that are determined by $\widehat{r}$) via shrinkage priors. An adaptive r-update step is further incorporated to improve the clustering performance. Simulations show that the proposed algorithms (3.2) and (3.3) have reasonable performance. Comparing to the conventional Dirichlet process modeling, it yields better posterior contraction, but at the cost of reduced empirical coverage. Further theoretical investigations are necessary to understand their asymptotic behaviors. In practice, this idea can be generalized to clustering problem of multi-dimensional data as well, i.e., $y_i = \theta_i \in \mathbb{R}^d$. Since we can not rank multivariate vectors, the possible alternative is to construct a

minimum spanning tree (MST) (Li and Sang, 2018), and impose shrinkage fusion prior on the pair of θ_i that is connected by a edge in the MST.

Acknowledgments. This work is in memory of Prof. Jayanta Ghosh who has jointly supervised the first PhD student of the second author. Qifan Song's research is sponsored by NSF DMS-1811812. Guang Cheng's research is sponsored by NSF DMS-1712907, DMS-1811812, DMS-1821183, and Office of Naval Research, (ONR N00014-18-2759).

References

ANDREWS, D.F. and MALLOWS, C.L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B (Methodological)*, 99–102.

BARRON, A. (1998). Information-theoretic characterization of bayes performance and the choice of priors in parametric and nonparametric problems. In J.M. Bernardo, J. Berger, A. Dawid, A. Smith, eds. *Bayesian Statistics 6*, 27–52.

BERGER, J.O., WANG, X. and SHEN, L. (2014). A bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics* **24**, 1, 110–129.

BETANCOURT, B., RODRÍGUEZ, A. and BOYD, N. (2017). Bayesian fused lasso regression for dynamic binary networks. *Journal of Computational and Graphical Statistics* **26**, 4, 840–850.

BHATTACHARYA, A., PATI, D., PILLAI, N.S. and DUNSON, D.B. (2015). Dirichlet-laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110**, 1479–1490.

CARVALHO, C.M., POLSON, N.G. and SCOTT, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40**, 4, 2069–2101.

CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A.W. (2015). Bayesian linear regression with sparse priors. *Annals of Statistics*, 1986–2018.

CHEN, J. and CHEN, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771.

CHEN, J. and CHEN, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica* **22**, 555–574.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

GHOSAL, S., GHOSH, J.K. and VAN DER VAART, A.W. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28**, 2, 500–531.

GHOSAL, SUBHASHIS and VAN DER VAART, A.W. (2007). Convergence rates of posterior distributions for noniid observations. *Annals of Statistics* **35**, 1, 192–223.

HAHN, P.R. and CARVALHO, C.M. (2015). Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association* **110**, 435–448.

HELLER, K.A. and GHAHRAMANI, Z. (2005). Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, 297–304.

ISHWARAN, H. and RAO, J.S. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, 730–773.

JIANG, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rate of the fitted densities. *Annals of Statistics* **35**, 1487–1511.

JOHNSON, V.E. and ROSSEL, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* **107**, 649–660.

JOHNSTONE, I.M. (2010). High dimensional bernstein-von mises: simple examples. *Institute of Mathematical Statistics Collections* **6**, 87.

KE, Z.T., FAN, J. and WU, Y. (2015a). Homogeneity pursuit. *Journal of the American Statistical Association* **110**, 509, 175–194.

KE, Z.T., FAN, J. and WU, Y. (2015b). Homogeneity pursuit. *Journal of the American Statistical Association* **110**, 175–194.

KLEIJN, B.J.K., VAN DER VAART, A.W. et al. (2006a). Misspecification in infinite-dimensional bayesian statistics. *The Annals of Statistics* **34**, 2, 837–877.

KLEIJN, B.J.K. and VAN DER VAART, A.W. (2006b). Misspecification in infinite-dimensional bayesian statistics. *Annals of Statistics* **34**, 837–877.

KYUNG, M., GILL, J., GHOSH, M. and CASELLA, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis* **5**, 2, 369–411.

LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 1302–1338.

LI, H. and PATI, D. (2017). Variable selection using shrinkage priors. *Computational Statistics & Data Analysis* **107**, 107–119.

LI, FURONG and SANG, HUIYAN (2018). Spatial homogeneity pursuit of regression coefficients for large datasets. *Journal of the American Statistical Association, (just-accepted)*, 1–37.

LIANG, F., SONG, Q. and YU, K. (2013). Bayesian subset modeling for high dimensional generalized linear models. *Journal of the American Statistical Association* **108**, 589–606.

LIU, J., YUAN, L. and YE, J. (2010). An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 323–332.

MA, S. and HUANG, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association* **112**, 517, 410–423.

MOZEIKA, A. and COOLEN, A. (2018). Mean-field theory of bayesian clustering. *arXiv*:1709.01632.

NARISETTY, N.N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* **42**, 2, 789–817.

NEAL, R.M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics* **9**, 2, 249–265.

PARK, T. and CASELLA, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.

RINALDO, A. et al. (2009). Properties and refinements of the fused lasso. *The Annals of Statistics* **37**, 5B, 2922–2952.

ROBBINS, H. (1985). An empirical bayes approach to statistics. In *Herbert Robbins Selected Papers*, 41–47.

ROYSTON, J.P. (1982). Algorithm as 177: Expected normal order statistics (exact and approximate). *Journal of the Royal Statistical Society. Series C (Applied statistics)* **31**, 2, 161–165.

SCOTT, J.G. and BERGER, J.O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 2587–2619.

SHEN, X. and HUANG, H.-C. (2012). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association* **105**, 727–739.

SHIMAMURA, K., UEKI, M., KAWANO, S. and KONISHI, S. (2018). Bayesian generalized fused lasso modeling via neg distribution. *Communications in Statistics-Theory and Methods*, 1–23.

SONG, Q. and LIANG, F. (2014). A split-and-merge bayesian variable selection approach for ultra-high dimensional regression. *Journal of the Royal Statistical Society, Series B*, in press.

SONG, Q. and LIANG, F. (2017). Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv*:1712.08964.

TANG, X., XU, X., GHOSH, M. and GHOSH, P. (2016). Bayesian variable selection and estimation based on global-local shrinkage priors. *arXiv*:1605.07981.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., JI, Z. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 1, 91–108.

TIBSHIRANI, R. and WANG, P. (2007). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics* **9**, 1, 18–29.

VAN DER GEER, S. and BÜHLMANN, P. (2011). *Statistics for High-Dimensional Data Methods, Theory and Applications*. Spring Series in Statistics, Springer.

VAN DER PAS, S.L., SZABO, B. and VAN DER VAART, A. (2017). Adaptive posterior contraction rates for the horseshoe. *arXiv*:1702.03698.

WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis* **13**, 559–626.

XU, Z., SCHMIDT, D.F., MAKALIC, E., QIAN, G. and HOPPER, J.L. (2017). Bayesian sparse global-local shrinkage regression for grouped variables. arXiv:1709.04333.

YANG, Y., WAINWRIGHT, M.J. and JORDAN, M.I. (2015). On the computational complexity of high-dimensional bayesian variable selection. *Annals of Statistics*, in press.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894–942.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

ZUBKOV, A.M. and SEROV, A.A. (2013). A complete proof of universal inequalities for the distribution function of the binomial law. *Theory of Probability & Its Applications* **57**, 539–544.

A. Appendix

First, let us state some useful lemmas.

Lemma A.1 (Lemma 1 of Laurent and Massart 2000)**.** *Let $\chi^2_d(\kappa)$ be a chi-square distribution with degree of freedom d, and noncentral parameter κ, then we have the following concentration inequality*

$$Pr(\chi^2_d(\kappa) > d + \kappa + 2x + \sqrt{(4d + 8\kappa)x}) \leq \exp(-x), \text{ and}$$

$$Pr(\chi^2_d(\kappa) < d + \kappa - \sqrt{(4d + 8\kappa)x}) \leq \exp(-x).$$

Lemma A.2 (Theorem 1 of Zubkov and Serov 2013)**.** *Let X be a Binomial random variable $X \sim \mathrm{B}(n, v)$. For any $1 < k < n - 1$*

$$Pr(X \geq k + 1) \leq 1 - \Phi(sign(k - nv)\{2nH(v, k/n)\}^{1/2}),$$

where Φ is the cumulative distribution function of standard Gaussian distribution and $H(v, k/n) = (k/n)\log(k/nv) + (1 - k/n)\log[(1 - k/n)/(1 - v)]$.

The next lemma is a refined result of Lemma 6 in Barron (1998):

Lemma A.3. *Let f^* be the true probability density of data generation, f_θ be the likelihood function with parameter $\theta \in \Theta$, and E^*, E_θ denote the corresponding expectation respectively. Let B_n and C_n be two subsets in parameter space Θ, and ϕ_n be some test function satisfying $\phi_n(D_n) \in [0, 1]$ for any data D_n. If $\pi(B_n) \leq b_n$, $E^*\phi(D_n) \leq b'_n$, $\sup_{\theta \in C_n} E_\theta(1-\phi(D_n)) \leq c_n$, and furthermore,*

$$P^*\left\{\frac{m(D_n)}{f^*(D_n)} \geq a_n\right\} \geq 1 - a'_n,$$

where $m(D_n) = \int_\Theta \pi(\theta) f_\theta(D_n) d\theta$ is the margin probability of D_n. Then,

$$E^*\left(\pi(C_n \cup B_n)|D_n)\right) \leq \frac{b_n + c_n}{a_n} + a'_n + b'_n.$$

Proof. Define Ω_n to be the event of $(m(D_n))/(f^*(D_n)) \geq a_n$, and $m(D_n, C_n \cup B_n) = \int_{C_n \cup B_n} \pi(\theta) f_\theta(D_n) d\theta$. Then

$$\begin{aligned}
& E^*\pi(C_n \cup B_n)|D_n) = E^*\pi(C_n \cup B_n)|D_n)(1 - \phi(D_n))1_{\Omega_n} \\
+ \;& E^*\pi(C_n \cup B_n)|D_n)(1 - \phi(D_n))(1 - 1_{\Omega_n}) + E^*\pi(C_n \cup B_n)|D_n)\phi(D_n) \\
\leq \;& E^*\pi(C_n \cup B_n)|D_n)(1 - \phi(D_n))1_{\Omega_n} + E^*(1 - 1_{\Omega_n}) + E^*\phi(D_n) \\
\leq \;& E^*\pi(C_n \cup B_n)|D_n)(1 - \phi(D_n))1_{\Omega_n} + b'_n + a'_n \\
\leq \;& E^*\{m(D_n, C_n \cup B_n)/a_n f^*(D_n)\}(1 - \phi(D_n)) + b'_n + a'_n.
\end{aligned}$$

By Fubini theorem,

$$\begin{aligned}
& E^*(1 - \phi(D_n))m(D_n, C_n \cup B_n)/f^*(D_n) = \int_{C_n \cup B_n} \int_{\mathcal{X}} [1 - \phi(D_n)] f_\theta(D_n) dD_n \pi(\theta) d\theta \\
\leq \;& \int_{C_n} E_\theta(1 - \phi(D_n))\pi(\theta) d\theta + \int_{B_n} \int_{\mathcal{X}} f_\theta(D_n) dD_n \pi(\theta) d\theta \leq b_n + c_n.
\end{aligned}$$

Combining the above inequalities leads to the conclusion.

PROOF OF THEOREM 2.1 AND 2.2.

Let $G = \{g_1, g_2, \ldots, g_d\}$ be a generic subset of $\{2, \ldots, n\}$, and it also represents potential a $(d+1)$-group structure of θ as $\{\{1, \ldots, g_1\}, \{g_1 + 1, \ldots, g_2\}, \ldots, \{g_d + 1, \ldots, n\}\}$. Given G and its corresponding blocking structure, $\widehat{\theta}_G(y)$ denotes the estimation of θ based on block mean, i.e. $\widehat{\theta}_{G,j}(y) = \sum_{i=g_j+1}^{g_{j+1}} y_i/(g_{j+1}-g_j)$ for all $g_j+1 \le j \le g_{j+1}$, and $\widehat{\sigma}^2_G(y) = \|y-\widehat{\theta}_G\|^2/(n-|G|-1)$.

To prove the posterior contraction, we will apply Lemma A.3. We define the following testing function

$$\phi(y) = 1\{ \quad \|\widehat{\theta}_G - \theta^*\| \ge \sqrt{n}\sigma^*\epsilon_n \text{ and } |\widehat{\sigma}^2_G - \sigma^{*2}| > \sigma^{*2}\epsilon_n \text{ for all } G \supset G^*, |G| \le (1+\delta)|G^*|\} \tag{A.1}$$

for some $\delta > 0$, and define C_n and B_n as:

$$C_n = \{\theta : \|\theta - \theta^*\| \le M\sqrt{n}\sigma^*\epsilon_n, (1-\epsilon_n)/(1+\epsilon_n) < \sigma^2/\sigma^{*2} < (1+\epsilon_n)/(1-\epsilon_n)\}^c \backslash B_n,$$
$$B_n = \{\theta : \text{Among all } \{\vartheta_i' s\}_{i \notin G^*}, \text{ there are at least } \delta|G^*| \text{ of them are greater than } \sigma\epsilon_n/n\}.$$

Note that when $G \supset G^*$, $\|\widehat{\theta}_G(y) - \theta^*\|^2 \sim \sigma^{*2}\chi^2_{|G|+1}$ and $\|y - \widehat{\theta}_G\|^2 \sim \sigma^{*2}\chi^2_{n-|G|-1}$, thus by Lemma A.1, we have that

$$P(\|\widehat{\theta}_G(y) - \theta^*\| \ge \sqrt{n}\sigma^*\epsilon_n \text{ and } |\widehat{\sigma}^2_G - \sigma^{*2}| > \sigma^{*2}\epsilon_n) \le \exp\{-c_1 n\epsilon_n^2\},$$

for some constant c_1, since $|G| = O(|G^*|) \prec n\epsilon_n^2$ and $\epsilon_n \prec 1$. Therefore,

$$E_{(\theta^*,\sigma^{*2})}\phi(y) \le \binom{n-1}{(1+\delta)|G^*|} \exp\{-c_1 n\epsilon_n^2\} = \exp\{-c_1' n\epsilon_n^2\}, \tag{A.2}$$

as long as $n\epsilon^2/[|G^*|\log n]$ is sufficiently large.

For any $(\theta, \sigma^2) \in C_n$ satisfying $\|\theta - \theta^*\| \le M\sqrt{n}\sigma^*\epsilon_n$ and $\sigma^2/\sigma^{*2} \ge (1-\epsilon_n)/(1+\epsilon_n)$, we define $\widehat{G} = \{i : \theta_i - \theta_{i-1} \ge \sigma\epsilon_n/n^2\} \cup G^*$ (hence $|G| \le (1+\delta)|G^*|$), thus

$$\begin{aligned}
& P_{(\theta,\sigma^2)}(\|\widehat{\theta}_{\widehat{G}}(y) - \theta^*\| \le \sqrt{n}\sigma^*\epsilon_n) = P_{(\theta,\sigma^2)}(\|\widehat{\theta}_{\widehat{G}}(y) - \widehat{\theta}_{\widehat{G}}(\theta) + \widehat{\theta}_{\widehat{G}}(\theta) - \theta^*\| \le \sqrt{n}\sigma^*\epsilon_n) \\
\le\ & P_{(\theta,\sigma^2)}(\|\widehat{\theta}_{\widehat{G}}(y) - \widehat{\theta}_{\widehat{G}}(\theta)\| \ge \|\widehat{\theta}_{\widehat{G}}(\theta) - \theta^*\| - \sqrt{n}\sigma^*\epsilon_n) \\
\le\ & P_{(\theta,\sigma^2)}(\|\widehat{\theta}_{\widehat{G}}(y) - \widehat{\theta}_{\widehat{G}}(\theta)\| \ge M\sqrt{n}\sigma^*\epsilon_n - \sqrt{n}\sigma\epsilon_n - \sqrt{n}\sigma^*\epsilon_n) \\
\le\ & P\left(\chi^2_{|G+1|} \ge \left[\sqrt{\frac{1-\epsilon_n}{1+\epsilon_n}}(M-1) - 1\right] n\epsilon_n^2\right) \le \exp\{-c_2 n\epsilon_n^2\}
\end{aligned}$$

for some c_2 given a large M, where the second inequality is due to the fact that $\|\widehat{\theta}_{\widehat{G}}(\theta) - \theta\| \leq \sqrt{n}\sigma\epsilon_n$ when $\theta \in B_n$.

For any $(\theta, \sigma^2) \in C_n$ satisfying $\sigma^2/\sigma^{*2} < (1-\epsilon_n)/(1+\epsilon_n)$ or $\sigma^2/\sigma^{*2} > (1+\epsilon_n)/(1-\epsilon_n)$,

$$\begin{aligned}
& P_{(\theta,\sigma^2)}(|\widehat{\sigma}^2_G - \sigma^{*2}| < \sigma^{*2}\epsilon_n) = P_{(\theta,\sigma^2)}(|\|y - \widehat{\theta}_G\|^2/\sigma^{*2}(n - |G| - 1) - 1| < \epsilon_n) \\
\leq\ & P_{(\theta,\sigma^2)}(1 - \epsilon_n < \|y - \widehat{\theta}_G\|^2/\sigma^{*2}(n - |G| - 1) < 1 + \epsilon_n) \\
\leq\ & P_{(\theta,\sigma^2)}\left(\left|\frac{\|y - \widehat{\theta}_G(y)\|^2}{\sigma^2} - (n - |G| - 1)\right| > (n - |G| - 1)\epsilon_n\right) \\
=\ & P_{(\theta,\sigma^2)}\left(|\chi^2_{n-|G|-1}(\lambda) - (n - |G| - 1)| > (n - |G| - 1)\epsilon_n\right) \leq \exp\{-c_2' n\epsilon_n^2\}
\end{aligned}$$

for some c_2', where the noncentral parameter $\lambda < n\epsilon_n^2 \prec (n - |G| - 1)\epsilon_n$.

Combining the results from the previous two paragraph, it is easy to obtain that

$$\sup_{(\theta,\sigma^2)\in C_n} E_{(\theta,\sigma^2)}[1 - \phi(y)] \leq \max\{\exp(-c_2 n\epsilon_n^2), \exp(-c_2' n\epsilon_n^2)\}. \tag{A.3}$$

Now we consider the marginal posterior density of data y. With probability $P(\|\varepsilon\| \leq 2\sqrt{n}\sigma^*)$ (which converges to 1),

$$\begin{aligned}
& \frac{m(y)}{f^*(y)} = \frac{\int_{\sigma^2}\int_\theta \sigma^{*n}\exp\{-\|\theta^* - \theta + \varepsilon\|^2/\sigma^2\}d\theta d\sigma^2}{\sigma^n \exp\{-\|\varepsilon\|^2/\sigma^{*2}\}} \\
\geq\ & \int_{\sigma^2}\int_\theta \exp\left\{-\frac{\|\theta^* - \theta\|^2}{\sigma^2} - 2\frac{(\theta^* - \theta)^T\varepsilon}{\sigma^2} + \frac{\|\varepsilon\|^2}{\sigma^{*2}} - \frac{\|\varepsilon\|^2}{\sigma^2} - n\log(\sigma/\sigma^*)\right\} \\
& \pi(\theta, \sigma^2)d\theta d\sigma^2 \\
\geq\ & \pi(\max\{|\theta_1 - \theta_1^*|, |\vartheta_i - \vartheta_i^*|\}/\sigma \leq |G^*|\log n/n^2, 0 \leq \sigma^2 - \sigma^{*2} \\
\leq\ & \sigma^{*2}|G^*|\log n/n)\exp\{-c_3'|G^*|\log n\}
\end{aligned}$$

for some constant c_3'. Besides,

$$\begin{aligned}
& \pi(\max\{|\theta_1 - \theta_1^*|, |\vartheta_i - \vartheta_i^*|\}/\sigma \leq |G^*|\log n/n^2, 0 \leq \sigma^2 - \sigma^{*2} \leq \sigma^{*2}|G^*|\log n/n) \\
\geq\ & \pi_\sigma(\sigma^{*2}) * O(\sigma^{*2}|G^*|\log n/n) * \underline{\pi}_\theta * O(|G^*|\log n/n^2) \\
& * \underline{\pi}_\vartheta^{|G^*|}[|G^*|\log n/n^2)]^{|G^*|} * [\pi_\vartheta(\{-|G^*|\log n/n^2, |G^*|\log n/n^2\})]^n \\
=\ & \exp\{-c_3''|G^*|\log n\}. \quad \text{(by the conditions imposed on the prior specifications)}
\end{aligned}$$

for some constant c_3''. Thus

$$m(y)/f^*(y) \geq \exp\{-(c_3' + c_3'')|G^*|\log n\} = \exp\{-c_3 n\epsilon_n^2\}, \text{ with probability tending to 1} \tag{A.4}$$

for some c_3, where c_3 can be sufficiently small when $n\epsilon_n^2/[|G^*|\log n]$ is large enough.

At last, we study the prior probability of set B_n. Due to the prior independence of ϑ_i's, $\pi(B_n) = \pi[\text{Bin}(n-1-|G^*|, p) > (\delta)|G^*|]$, where $p \le (1/n)^{1+u}$. By Lemma A.2,

$$\pi(B_n) \le \exp\{-c_4\delta|G^*|\log n\} \tag{A.5}$$

for some c_4. Combine results (A.2), (A.3), (A.4) and (A.5), and we apply Lemma A.3 to get the posterior consistency result that

$$\pi(B_n \cup C_n|y) \to^p 0,$$

given a sufficient large constants δ and $n\epsilon_n/|G^*|\log n$.

PROOF OF THEOREM 3.1.

Consider $y = \theta + d$, where $\theta_i^* \equiv 0$ for all i and error d is order statistics of standard normal variables, i.e. the density of d is $f(d) = n!\prod\phi(d_i)1(d_1 \le d_2, \ldots, d_{n-1} \le d_n)$ and ϕ denotes the standard normal density. The prior of θ follows $\pi(\theta) = \pi_1(\theta_1)\prod_{i=2}^n \pi_{t,s}(\theta_i - \theta_{i-1})$, where π_{λ_1} is the density of $N(0,\lambda_1)$, and $\pi_{t,s}$ is the density of t distribution with tiny scale parameter satisfying $-\log s \asymp \log n$, i.e. conditions in Corollary 2.1 holds, and we consider the misspecified posterior of form $\pi(\theta|D_n) = \exp\{-(y-\theta)^2/2\}\pi(\theta)$.

Define $\mu \in \mathbb{R}^n$ as $\mu_i = 0$ for all $1 \le i \le k = 3n/4$, and $\mu_i = Z_{0.25}/2$ for $i > k$ where $Z_{0.25}$ is the right 25% quantile of standard normal distribution, thus $\|\mu - \theta^*\|^2 \asymp n$.

Let $\Delta\theta$ be any vector such that $\|\Delta\theta\|^2 \le M\log n$. Then

$$-\log\left(\frac{\pi(\mu+\Delta\theta)}{\pi(\theta^*+\Delta\theta)}\right) = -\log\left(\frac{\pi_{t,s}(Z_{0.25}/2+\Delta\theta_k)}{\pi_{t,s}(\Delta\theta_k)}\right) = O(-\log s) = O(\log n).$$

And

$$\begin{aligned}
&\log\left(\frac{\exp\{-(y-\mu-\Delta\theta)^2/2\}}{\exp\{-(y-\theta^*-\Delta\theta)^2/2\}}\right) = [(y-\theta^*-\Delta\theta)^2 - (y-\mu-\Delta\theta)^2]/2 \\
= &\frac{1}{2}\sum_{i=k+1}^n [(y_i-\Delta\theta_i)^2 - (y_i - Z_{0.25}/2 - \Delta\theta_i)^2] = \frac{1}{2}\sum_{i=k+1}^n [(y_i-\Delta\theta_i)Z_{0.25} - \frac{Z_{0.25}^2}{4}] \\
\ge &\frac{1}{2}\left[(\|y_{k+1:n}\|_1 - \sqrt{nM\log n/4})Z_{0.25} - \frac{nZ_{0.25}^2}{16}\right] \ge cn,
\end{aligned}$$

for some positive constant c given sufficiently large n, where the inequalities above hold since $y_n \geq y_{n-1} \cdots \geq y_{k+1}$, and $y_{k+1} \approx Z_{0.25}$ with high probability, due to large sample empirical quantile theory.

Combining the above two results, we have that the posterior density satisfies $\pi(\mu + \Delta\theta|D_n) \gg \pi(\theta^* + \Delta\theta|D_n)$ for any $\|\Delta\theta\|^2 \leq M \log n$ with high probability. Therefore, more posterior mass is distributed within the $\sqrt{M \log n}$-radius ball centered at μ than at the true parameter θ^*.

Proof of Theorem 3.2.

The proof of this theorem is quite similar to the proof of Theorem 2.1 and 2.2. We define the same testing function as in the proof of Theorem 2.1 and 2.2, and define the following two sets:

$$C_n = \{\theta : \|\theta - \theta^*\| \leq M\sqrt{n}\sigma^*\epsilon_n, (1-\epsilon_n)/(1+\epsilon_n) < \sigma^2/\sigma^{*2} < (1+\epsilon_n)/(1-\epsilon_n)\}^c \backslash B_n,$$
$$B_n = \{\theta : \text{Among all } \{\theta_i - \theta_{i-1}\}_{i=2}^n, \text{ there are at least} \delta \text{ of them are greater than } \sigma\epsilon_n/n\}.$$

Using the same arguments, one can still establish exponential separation results (A.2) and (A.3).

To establish (A.4), we notice that

$$\begin{aligned}
&\frac{m(y)}{f^*(y)} = \frac{\int_{\sigma^2}\int_\theta \sigma^{*n}\exp\{-\|\theta^* - \theta + \varepsilon\|^2/\sigma^2\}d\theta d\sigma^2}{\sigma^n \exp\{-\|\varepsilon\|^2/\sigma^{*2}\}} \\
\geq\ & \int_{\sigma^2}\int_\theta \exp\left\{-\frac{\|\theta^* - \theta\|^2}{\sigma^2} - 2\frac{(\theta^* - \theta)^T\varepsilon}{\sigma^2} + \frac{\|\varepsilon\|^2}{\sigma^{*2}} - \frac{\|\varepsilon\|^2}{\sigma^2} - n\log(\sigma/\sigma^*)\right\}\pi(\theta, \sigma^2)d\theta d\sigma^2 \\
\geq\ & \pi(\max\{|\theta_i - \theta_i^*|\}/\sigma \leq \log n/n, 0 \leq \sigma^2 - \sigma^{*2} \leq \sigma^{*2}\log n/n)\exp\{-c_3'\log n\}
\end{aligned}$$

for some constant c_3' and

$$\begin{aligned}
&\pi(\max\{|\theta_i - \theta_i^*|\}/\sigma \leq \log n/n, 0 \leq \sigma^2 - \sigma^{*2} \leq \sigma^{*2}\log n/n) \\
\geq\ & \sum_r \pi(\max\{|\theta_{r(1)} - \theta_{r(1)}^*|, |\theta_{r(i)} - \theta_{r(i-1)}|\}/\sigma \leq |G^*|\log n/n^2, 0 \\
\leq\ & \sigma^2 - \sigma^{*2} \leq \sigma^{*2}\log n/n|r)\pi(r).
\end{aligned}$$

This ensures (A.4).

As for the prior probability of B_n, if the scale parameter for the t distribution is sufficiently small, i.e. $s = n^{-w}$ for some large w and $\int_{\pm\epsilon_n/n^2} \pi_{t,s}(x)dx \geq$

$1-1/n^{1+u}$ for some sufficiently large u where $\pi_{t,s}$ denotes the t density function with scale parameter s, then for any ranking r,

$$\pi(B_n|r) \geq 1-\pi(\max\{\theta_{r(i)}-\theta_{r(i-1)}\} \leq \sigma\epsilon_n/n^2|r) \geq 1-(1-1/n^{1+u})^n \approx n^{-u}.$$

This hence implies that $-\log(\pi(B_n)) \geq u\log n$.

Qifan Song
Guang Cheng
Department of Statistics,
Purdue University,
610 Purdue Mall,
West Lafayette, IN,
47907, USA
E-mail: qfsong@purdue.edu
chengg@purdue.edu

Paper received: 26 December 2018.