Bayesian Fusion Estimation via t-Shrinkage

Qifan Song and Guang Cheng *

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 simulation shows that the proposed algorithm obtains better posterior distributional convergence than the classical Dirichlet process modeling.

Keywords: 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 [47] 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

*Qifan Song is an assistant professor in Department of Statistics, Purdue University. Email: qfsong@purdue.edu; Guang Cheng is a professor in Department of Statistics, Purdue University. Email: chengg@purdue.edu.
or using informative prior specifications. Various penalty functions have been proposed for consistent frequentist estimation, including Lasso [44], SCAD [11], adaptive Lasso [53] and MCP [52]. For high dimensional Bayesian inferences, sparsity induced prior, such as spike-and-slab prior [7, 14, 18, 28, 32, 38, 41, 51], is widely used for model selection. Fused Lasso [45] 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 \( \theta_i \)'s, such that \( \theta_i \)'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 \( \theta_i \)'s and \( \vartheta_i \)'s respectively.

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

\[
y_i = \theta^*_i + \varepsilon_i
\]

where \( \varepsilon_i \)'s are iid normal error with unknown variance \( \sigma^2 \). Similarly to Fused Lasso applications, we also assume true parameter \( \theta^* \) is blocky in the sense that there exists a partition \( \{ B^*_1, \ldots, B^*_s \} \) of \( \{1,...,n\} \) such that \( \theta^*_i \)'s are constant for all \( i \in B^*_k \). 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 \( \theta^* \). Motivated by the \( L^1 \)-fusion penalty used by Fused lasso estimator [45], as well as the development of the Bayesian lasso [34, 24] 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 comparing 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 [5, 7, 42]. It is believed that the posterior inconsistency of Laplace prior is due to its exponentially light tail, and [42] 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 \( \theta_i \) estimation rather than a blocky \( \theta \), 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 [40] 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 $\theta^*$ has a consecutive blocking structure, it is more realistic to assume that $\theta^*$ possesses an unknown clustering structure. In other words, some unnecessarily consecutive $\theta^*_i$’s share the same value within an $B^*_i$. In a broader scope, such a clustering problem can viewed as a simplest example of subgroup analysis where we assume that a subject-related parameter $\theta$ 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 subgroup. 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 Bayesian clustering analysis [3, 15, 31, 49] usually impose discrete priors on the clustering structure, along with a conditional prior on $\theta$ given specific clustering structure. In contrast, we propose to directly model the parameter $\theta$ via $t$ fusion prior. Such a prior specification allows a computationally efficient Gibbs posterior sampling algorithm. Our simulation shows that our procedure yields reasonable cluster structure recovery, and moreover it beats the usual Dirichlet process prior in terms of posterior contraction.

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. At last, Section 5 provides more discussions and remarks. All technical proofs are postponed to the appendix section.

Throughout this work, the following notation is used. Given two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \succ b_n$ means $\lim(a_n/b_n) = \infty$ and $a_n \sim b_n$ means $-\infty < \lim \inf(a_n/b_n) \leq \lim \sup(a_n/b_n) < \infty$. $\|x\|$
and $\|x\|_1$ denote $L_2$ and $L_1$ norms of vector $x$.

## 2 Bayesian fusion via $t$-shrinkage

### 2.1 Bayesian Modeling

Suppose we observe independent data $\{y_i : i = 1, \ldots, n\}$ following model (1). The indexing of the data has certain practical or scientific meaning, under which we can assume that the parameter vector $\theta^*$ 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 $\theta_i$'s and $\vartheta_i$'s, [45] proposed the following fused lasso estimator

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

If one is not interested in pursuing the sparsity of $\theta$'s, then a fusion estimator [35] can be used

$$\hat{\theta}_F = \text{arg min} \left( \frac{\|y - \theta\|^2}{2} + \lambda \sum_{i=2}^{n} |\vartheta_i| \right) = \text{arg min} \left( \frac{\|y - \theta\|^2}{2} + \lambda \sum_{i=2}^{n} |\theta_i - \theta_{i-1}| \right),$$

for some tuning parameter $\lambda$. The above objective functions are both convex and fast computation algorithms are developed, e.g. [29, 45]. The penalty term $\lambda \sum_{i=2}^{n} |\theta_i - \theta_{i-1}|$ is can be interpreted as the negative logarithm of prior density used for Bayesian inferences, therefore, a nature Bayesian expansion to (2) is Laplace (double exponential) prior modeling [24, 40]. To account for the unknown variance parameter $\sigma^2$ and $\theta_1$, a convenient prior specification could be

$$\sigma^2 \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma), \quad \theta_1 | \sigma^2 \sim N(0, \sigma^2 \lambda_1),$$

$$(\theta_i - \theta_{i-1}) | \sigma^2 \sim \text{Laplace}(\lambda/\sigma), \text{ for all } i = 2, \ldots, n$$

where Laplace($a$) denotes the distribution with cdf $f(x) \propto \exp(-a|x|)$.

According to [1], the above Laplace prior Laplace($\lambda/\sigma$) 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 [34]:

\[
\lambda_i^{-1} \sim \text{Inverse-Gaussian}(\lambda \sigma / |\theta_i - \theta_{i-1}|, \lambda^2) \text{ for all } i = 2, \ldots, n,
\]

\[
\sigma^2 \sim \text{Inverse-Gamma} \left( a_\sigma + n, b_\sigma + \frac{||y - \theta||^2}{2} + \frac{\theta_1^2}{2\lambda_1} + \sum_{i=2}^{n} \frac{(\theta_i - \theta_{i-1})^2}{2\lambda_i} \right),
\]

\[
\theta_i \sim N(\mu_i, \nu_i),
\]

where \(\nu_i^{-1} = 1/\sigma^2 + 1/\lambda_{i+1}\sigma^2 + 1/\lambda_i\sigma^2\) and \(\mu_i = \nu_i(y_i/\sigma^2 + \theta_{i+1}/\lambda_{i+1}\sigma^2 + \theta_{i-1}/\lambda_i\sigma^2)\), \(\lambda_{n+1}\) and \(\lambda_0\) are considered to be infinite, and 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 [5, 7, 42] point out that, if we impose independent \(\theta_i | \sigma^2 \sim \text{Laplace}(\lambda/\sigma)\) in high dimensional sparse regression under Laplace priors, then the induced posterior has only a sub-optimal contraction rate, or even diverges. In other words, the posterior distribution of \(\theta\) doesn’t contract into a small neighborhood around true value \(\theta^*\) appropriately. For a blocky parameter \(\theta\), 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 \(\theta\) doesn’t have a step-wise pattern at all.

Following the theoretical discovery of [42], we consider using a class of heavy tailed priors for the successive differences \(\vartheta_i\)'s. Specifically, this work will assign an \(t\)-shrinkage prior:

\[
\sigma^2 \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma), \quad \theta_1 | \sigma^2 \sim N(0, \sigma^2 \lambda_1),
\]

\[
(\theta_i - \theta_{i-1}) | \sigma^2 \sim t_{df}(s\sigma), \text{ for all } i = 2, \ldots, n
\]

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 update for \(\theta_i\)'s and \(\sigma^2\) are exactly the same as in (4) and the update of \(\lambda_i\)'s follows

\[
\lambda_i \sim \text{Inverse-Gamma} \left( a_t + 1/2, b_t + \frac{(\theta_i - \theta_{i-1})^2}{2\sigma^2} \right) \text{ for all } i = 2, \ldots, n.
\]
To further understand the difference between the Laplace fusion prior and t fusion prior, we compare their conditional prior \( \pi(\theta_i | \theta_{i-1}, \theta_{i+1}, \sigma) \). 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 to see that the conditional t-fusion prior allocates most of its prior mass at the two small neighborhoods centered at \( \theta_{i-1} \) and \( \theta_{i+1} \), given a sufficiently small scale parameter \( s \). In other word, the prior introduces a strong shrinkage effect on \( \theta_i \), towards either \( \theta_{i-1} \) or \( \theta_{i+1} \). Therefore, for all \( i = 2, \ldots, n \), \( \theta_i \) will merge with either \( \theta_{i-1} \) or \( \theta_{i+1} \) 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 \( \theta_i \) to be grouped with either \( \theta_{i-1} \) or \( \theta_{i+1} \). It is worth to mention that the NEG-fusion prior[40] also has a similar plot 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[40] only use this MAP for Bayesian inferences rather than the whole posterior distributional information. But the t prior is continuous differentiable everywhere (the functions displayed in the upper plot of Figure 1 is 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 induced by the t-fusion prior specification[5]. Our theoretical investigation follows the framework of[12], which studies the posterior convergence rate of coefficient \( \beta \) in high dimensional sparse regression models \( y = X \beta + \varepsilon \). Note the model[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 \( \beta = (\theta_1, \theta_2, \ldots, \theta_n) \) is an unknown sparse vector. The following theorem studies the general posterior convergence properties given an independent prior over all \( \theta_i \)’s.

**Theorem 2.1** (Posterior consistency). Assume that \(|G^*| \prec n/\log(n)\), and prior specification follows that \( \sigma^2 \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma) \), \( \theta_1 \) and \( \theta_i \)’s are conditionally independent given \( \sigma^2 \) with prior density
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.

\begin{align*}
\pi(\theta_1, \vartheta_i|s) &\propto (1/\sigma)^n f_\theta(\theta_1/\sigma) \prod_{i=2}^n f_\vartheta(\vartheta_i/\sigma). \quad \text{Furthermore, if} \\
\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(\overline{\pi}_\vartheta) &= O(\log n), \quad \text{where } \overline{\pi}_\vartheta = \min_{|x| \leq \max_i |\vartheta_i^*|} f_\vartheta(x), \\
-\log(\overline{\pi}_\theta) &= O(|G^*| \log n), \quad \text{where } \overline{\pi}_\theta = \min_{|x| \leq |\theta_1^*|} f_\theta(x), \\
&\quad a_\sigma \log(1/b_\sigma) + b_\sigma/\sigma^* + (a_\sigma + 2) \log(\sigma^*) = O(|G^*| \log n),
\end{align*}

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

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

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 [42], and for the sake of readability, the proof is provided in the Appendix. The first inequality of sufficient condition set (6) requires that the prior imposed on
\( \vartheta_i \)'s is highly concentrated around zero, such that it induces sufficient Bayesian shrinkage effect for those \( \vartheta_i \)'s whose true value is 0. The rest inequalities of (6) essentially require that the prior density at true parameter is at least of order \( e^{-cn_{\epsilon}^2} \) for some \( c \), and this helps to prevent over-shrinkage for those \( \vartheta_i \)'s whose true value is not 0. Similar conditions, which need the prior to be “thick” at true parameter values, are regularly used in Bayesian literature [12, 13, 17, 23]. Given the concrete forms for prior density \( f_0 \) and \( f_\vartheta \), the second and third inequalities of (6) are equivalent to some upper bound constraints on the magnitude of \( \theta_i^\ast \) and \( \max_i |\theta_i^\ast| \)'s (see e.g., Corollary 2.1). The fourth inequality of (6) trivially holds for any fixed \( a_\sigma \) and \( b_\sigma \) if \( \sigma^\ast_2 \) 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 [19] commonly assume that \( \sigma^\ast_2 \propto n^{-1} \), then one can choose a fixed \( a_\sigma \) and \( b_\sigma \approx n^{-\kappa} \) for some \( \kappa > 0 \). Under such a choice, condition (6) holds as long as \( -k \log n \leq \log(\sigma^\ast_2) < K \log n \) for some positive constant \( K \).

The above result states that almost all the posterior mass contract into a neighborhood of \( \theta^\ast \) with a radius \( M\sigma^\ast \sqrt{n_{\epsilon_n}} \), that is, the posterior convergence rate is of an order \( \sigma^\ast \sqrt{\log n} \). Note that if the partition index set \( G^\ast \) were known, the oracle rate of contraction turns out to be \( O(\sigma^\ast \sqrt{\log |G^\ast|}) \). Hence, the Bayesian shrinkage achieves the ideal risk up to a logarithmic term in \( n \). In frequentist literature, Theorem 2.7 of [35] showed that the convergence rate of Fused Lasso is no larger than \( O(\sigma^\ast \sqrt{\log |G^\ast|}) \). However, this rate is not directly comparable since an additional minimal signal strength condition, which ensures that \( G^\ast \) can be fully recovered in probability, is imposed in this paper.

The posterior distribution of \( \vartheta \) are always continuous, and doesn’t directly provide Bayesian inferences for the block structure, or equivalently, the unknown \( G^\ast \). The following result characterizes the some asymptotic performance of posterior block partition via discretization.

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

\[
\pi(\{|G(\theta, \sigma)\setminus G^\ast| > \delta|G^\ast|\}|y) \to 0,
\]

for some fixed constant \( \delta \), 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 \( (\theta, \sigma^2) \) and mapping \( G(\cdot, \cdot) \), can be treated as a discrete posterior distribution for the unknown \( G^\ast \) and used for
Bayesian block partition selection. Theorem 2.2 essentially claims that the number false positive selections of 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 [5]. It is worth to note that if certain minimal signal strength condition holds as well, i.e., $\min_{\theta_i > 0} |\theta_i^*|$ 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 [42]. Readers of interests can easily derive such posterior selection consistency by themselves.

It is not difficult to verify the imposed conditions for the proposed $t$ fusion shrinkage prior [5].

**Corollary 2.1.** When $f_\theta$ is the cdf of normal distribution $N(0, \lambda_1)$ and $f_\varphi$ is the cdf of a $t$-distribution with scale parameter $s$, then the first three inequalities of (6) hold when

$$\log(\max_i |\theta_i^*|/\sigma^*) = O(\log(n)),$$

$$\theta_i^2/\lambda_1\sigma^2 + \log(\lambda_1) = O(\log(n)),$$

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 $f_\varphi$, as long as $f_\varphi$ 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 $\max_i |\theta_i^*|/\sigma^* = o(1)$.

### 2.3 Bayesian posterior inference

In this section, we will illustrate the posterior behavior of Bayesian $t$-fusion by a toy example, and compare it to Bayesian Laplace-fusion. We will also discuss other issues related to hyperparameter choice.

A simulation data was generated with $n = 100$ and $\sigma^* = 0.5$. The data and underlying true parameter value are plotted in Figure 2. Three estimation procedures are considered: 1) $L_1$ fusion estimation [2] where the tuning parameter are selected by cross validation; 2) Bayesian Laplace prior [3] with $\lambda = \sqrt{2 \log(n)}$ and $\lambda_1 = 5$; 3) Bayesian $t$ prior [5] with $a_t = 2, b_t = 0.005, a_\sigma = 0.5, b_\sigma = 0.5$ and $\lambda_1 = 5$. The posterior samples are obtained by Gibbs sampler for 2000 iterations. The frequentist estimator and boxplots of Bayesian marginal posterior distributions are displayed in Figure 2. Note that
for the sake of readability of the figure, we don’t draw the outliers in these boxplots.

Figure 2: Upper Left: Simulated toy data; Upper right: Frequentist fusion estimation; Lower left: Marginal posterior boxplots for each $\theta_i$ under Laplace fusion prior; Lower right: Marginal posterior boxplots under $t$ fusion prior. The red curve denotes the true $\theta_i^*$’s which contain three blocks.

The comparison shows that the $L_1$ fusion penalty leads to a strictly sparse $\hat{\theta}$ estimation and the estimated $\hat{\theta}$ 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 $\theta_i$. Comparing 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 more close 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 has shrinkage effect for the successive differences, and hence the data fluctuation carries over to the posterior distribution of $\theta$. In this case, it is not clear to recover the underlying block structure.
Note that for both (3) and (5) prior specifications, their posterior distributions highly depend on the choice of hyperparameter, 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 Figure 3.

![posterior boxplot for Laplace prior](image1)
![posterior boxplot for $t$ prior](image2)

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

From Figure 3 and its comparison with Figure 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 $\theta_i$’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 Figure 2. The choice of scale parameter
only affects the smoothness of posterior of \( \theta \). Smaller scale parameter leads to smoother non-parametric estimation which is similar to smooth spline regression; larger scale parameter leads to a rather rugged estimation. This phenomenon can be partially explained by Figure 1. As the scale of the Laplace prior decreases (i.e., \( \lambda \) increases), the conditional prior of \( \theta_i \) 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 value \(|\theta_{i+1} - \theta_{i-1}|\) when the scale parameter decreases to zero, regardless of the observation value \(y_i\).

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 (6), 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 [9, 10], choosing a proper prior hyperparameter posts many difficulty for Bayesian statisticians. Conventional choices include imposing a hyper-prior on the hyperparameter [8, 15] or empirical Bayes [36]. For high dimensional sparse GLMs, [28] suggested choosing a hyperparameter such that the posterior mean and posterior mode are close. Several Bayesian works [4, 40] considered 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 hyperparameter, i.e., the scale parameter of the \(t\) prior. An empirical suggestion based on authors’ experience is to choose the scale parameter of (5) such that

\[
P(\|t_{df}(s)\| \geq \sqrt{\log(n)/n}) \approx 1/n. \tag{7}
\]

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 \(a_t = 2, b_t = 0.005\) approximately satisfies (7).
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 $\theta$ in model (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 only consists of 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 [3, 15]. The resultant posterior sampling hence usually requires an inefficiently reversible-jump Metropolis-Hastings move to travel across different clustering models in the sampling space. Another common approach is to model a mixture distribution via nonparametric priors such as Dirichlet process or Chinese restaurant process [33]. These mentioned approaches enforce explicit cluster 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 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 $\theta_i$ 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 $\hat{r}(\cdot)$. For model (1), a trivial estimator is the rank statistic of the responses $y_i$, i.e., $y_{\hat{r}(i)} = y_{(i)}$ where $y_{(i)}$ denotes the order statistics of $y_i$’s. Thus this leads to hybrid Bayesian modeling, depending on a frequentist pilot estimator $\hat{r}$:

$$\sigma^2 \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma), \quad \theta_{\hat{r}(1)}|\sigma^2 \sim N(0, \sigma^2 \lambda_1),$$

$$\theta_{r(i)} - \theta_{r(i-1)}|\sigma^2, \lambda_i \sim N(0, \sigma^2 \lambda_i), \quad \lambda_i \sim \text{Inverse-Gamma}(a_t, b_t) \quad \text{for all } i = 2, \ldots, n.$$  

(8)
The corresponding posterior Gibbs sampler follows:

\[
\lambda_i \sim \text{Inverse-Gamma} \left( a_t + \frac{1}{2}, b_t + \frac{(\hat{\theta}_{r(i)} - \hat{\theta}_{r(i-1)})^2}{2\sigma^2} \right) \text{ for all } i = 2, \ldots, n.
\]

\[
\sigma^2 \sim \text{Inverse-Gamma} \left( a_{\sigma} + n, b_{\sigma} + \frac{\|y - \theta\|^2}{2} + \frac{\hat{\theta}_{r(i)}^2}{2\lambda_i} + \frac{\sum_{i=2}^{n} (\hat{\theta}_{r(i)} - \hat{\theta}_{r(i-1)})^2}{2\lambda_i} \right),
\]

\[
\theta_{r(i)} \sim N(\mu_i, \nu_i),
\]

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

The idea of taking advantage of a pilot estimator \( \hat{\theta} \) and its rank statistic \( \hat{r}(i) \) has also been implemented in the two-stage frequentist approach [20].

It is worth to mention that posterior consistency Theorems 2.1 and 2.2 for Bayesian shrinkage fusion don’t apply to the above Bayesian modeling [5] even when \( \theta_{r(i)} \) is in a ascending order, because \( \varepsilon_{r(i)} \) are no longer iid error observations. Such a data dependent prior [5] can also be interpreted as a misspecified Bayesian modeling [22], 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 [8] is quite natural and intuitive, there are several issues. First, comparing 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 Figure 4 than the toy data showed in Figure 2. Secondly, [8] fails to take account of uncertainty of the estimator \( \hat{r} \), and the strict monotonicity of \( y_{r(i)} \) will always carry over to the posterior of \( \theta_{r(i)} \). 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 \( \theta_{r(i)} \)'s should contract towards the best step function, in the sense of minimizing the Kullback-Leibler divergence [22], 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 [8] is used (except that there is no necessity to impose a prior on \( \sigma^2 \)) and the scale parameter satisfies \(-\log(b_t) \preceq \log n\), then in high probability

\[
\pi(||\theta|| \leq \sqrt{M \log n} | y) < 1/2,
\]
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 $\hat{r}$, we update the “working” order $r$ of $\theta$ over iterations. To be specific, we modify the above Gibbs sampling iteration as:

update $\theta_{r(i)}$’s, $\lambda_i$’s and $\sigma^2$ according to (9),

update the $r(\cdot)$ to be the rank statistic of the current sample $\theta$. 

(10)

In other words, we update $r(\cdot)$ to the rank statistic of the newly obtained $\theta$.

There are a couple of rationales behind the algorithm (10). First, the update of $r$ potentially allows algorithm (10) to incorporate the uncertainty of rank statistic into the posterior sampling. Heuristically, if the current $r(\cdot)$ is close to the true ranking, (10) will shuffle the ordering of $\theta$ 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 (10) can be somehow connected with a full Bayesian modeling:

$r \sim \text{Uniform distribution over all possible permutations},$

$\sigma^2 \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma), \quad \theta_{r(1)}|\sigma^2 \sim N(0, \sigma^2 \lambda_1),$

$(\theta_{r(i)} - \theta_{r(i-1)})|\sigma^2, \lambda_i \sim N(0, \sigma^2 \lambda_i), \quad \lambda_i \sim \text{Inverse-Gamma}(a_i, b_i) \quad \text{for all } i = 2, \ldots, n.$

Under (11), the conditional posterior of $r(\cdot)$ follows $\pi(r|\theta, \lambda_i')s, y) \propto \exp\{-\sum_{i=2}^{n}(\theta_{r(i)} - \theta_{r(i-1)})^2/\lambda_i^2 \sigma^2\}$, from which sampling is expensive. However, when most of $\lambda_i^2$s 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 $\theta$. Therefore, updating $r$ to the rank statistic of current sample $\theta$, as in (10), 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. $\theta^*$ is a vector of $\theta_0$’s. If prior (11) is used with $b_t = n^{-c}$ for some sufficiently large $c$, and $\theta_0/(\lambda_1 \sigma^2) + \log(\lambda_1) = O(\log n)$ then there exist constant $M$ and $\epsilon_n \sim \sqrt{\log n/n}$, such that

$\pi(\|\theta - \theta^*\| \geq M \sigma^* \sqrt{\epsilon_n} | y) \rightarrow 0,$
where the convergence holds in probability and $L_1$.

Further investigations will be pursued to assess the posterior contraction induced by (11) when the true parameter $\theta^*$ has multiple-cluster structure. The minimal cluster difference $\min_{(i,j): \theta_i \neq \theta_j} |\theta_i - \theta_j|$ shall play an 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 (11) leads to satisfactory Bayesian consistency.

It is worth to mention that the $r$-update step in algorithm (10) may critically change the stochastic stability of the algorithm and that the ergodicity of Markov chains generated by (10) is still unclear. Therefore, in our application we use it with caution. In all our simulations and toy examples, we initialize with $r = \hat{r}$, and the $r$-update step is only implemented every other 20 iterations after certain burn-in period.

In frequentist literature, another popular approach to cluster $\theta_i$'s is to impose a pairwise difference penalty as $\sum_{i \neq j} p(|\theta_i - \theta_j|)$ for some penalty $p(\cdot)$\cite{21, 30, 39}. 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 $\pi$, 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 $\theta_i$'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 $\theta_i$ 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 (9) and (10), a simple toy example is conducted. We simulate a data set with $n = 100$ data points that belong to 3 underlying subgroups. For both (9) and (10), 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

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

(12)

and Bayesian Dirichlet process modeling

$$y_i|\theta_i, \sigma^2 \sim N(\theta_i, \sigma^2), \quad \theta_i \sim G,$$

$$G \sim \text{Dirichlet-Process}(N(0, \lambda), a) \text{ and } \sigma^2 \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma)$$  

(13)

with $\lambda = 5$, $a = 0.1$ and $a_\sigma = b_\sigma = 0.5$. The posterior sampling algorithms of Dirichlet process modeling are discussed in [33]. The frequentist estimator and posterior boxplots of $\theta_i$s under different priors are plotted in Figure 4. This figure clearly shows that frequentist $L_1$ fusion estimator fails to yield clustering structure for monotone sorted $y_{\hat{F}(i)}$’s, especially for the middle portion of the data, one always have $\hat{\theta}_{\hat{F}(i)} = y_{(i)}$. 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 [8] 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 occurs, and the posterior identifies more than 6 clusters. At last, the posterior obtained by algorithm [10] combines both advantages of DP prior and [8]. Comparing with DP prior, it has much smaller posterior variance; comparing with prior [8], the over-clustering issue is greatly alleviated.

4 Simulation and data anlaysis

4.1 Bayesian fusion simulations

In our first simulation studies, we consider model [1] with $\sigma^* = 0.5$ and true parameter $\theta^* \in \mathbb{R}^{100}$ having three consecutive blocks. These three blocks are $B^*_1 = \{1, \ldots, b_1\}$, $B^*_2 = \{b_1 + 1, \ldots, b_1 + b_2\}$, $B^*_3 = \{b_1 + b + 2 + 1, \ldots, b_1 + b_2 + b_3\}$ where $(b_1, b_2, b_3) \sim \text{multinomial}(100, (1/3, 1/3, 1/3))$, and $\theta^*_i = 0, 2$ or 4 within each block respectively. We compare the performance among Bayesian $t$ fusion [5], Bayesian Laplace fusion [3] and frequentist $L_1$ fusion [2], 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 $L_2$ and $L_1$ estimation errors, $\|\hat{\theta} - \theta^*\|_2$ and $\|\hat{\theta} - \theta^*\|_1$, 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 \( \Sigma \): \( R = \|\Sigma^\star - \hat{\Sigma}\|_1 = \sum_{i,j} |\hat{\omega}_{ij} - \omega_{ij}^\star| \). For \( L_1 \) fusion estimator, \( \hat{\Sigma} \) is trivially estimated by \( \hat{\omega}_{ij} = 1\{\hat{\theta}_i^F = \hat{\theta}_j^F\} \). 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 \( \Omega \). In the literature, this is usually done by 1) hard thresholding \([6, 16, 27, 43]\), or 2) decoupling shrinkage and selection methods \([14, 50]\). All these mentioned approaches depend solely on the posterior scaling, and don’t take into account of the degree of prior concentration. Therefore, following the suggestion made in \([32]\), we estimate \( \hat{\omega}_{ij} = 1\{\hat{\theta}_i/\hat{\sigma} = \hat{\theta}_j/\hat{\sigma} \leq 1/2n\} \), where \( \hat{\theta}_i \) and \( \hat{\sigma} \) are the Bayes estimator, \( \pi_{1/2n} \) denotes the \( 1 - 1/2n \) quantile of the prior distribution imposed on successive difference \( \hat{\theta}_i/\sigma \). 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, $\pi_{1/2n} = st_{1/2n}$ where s is the scale parameter, $t_{1/2n}$ is the $1 - 1/2n$ quantile of t distribution with df= $2\alpha$; for Laplace shrinkage, $\pi_{1/2n} = \log(n)/\lambda$. 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: $W = \text{Average}_{\{w_{ij} \neq 0\}} |\hat{\theta}_i - \hat{\theta}_j|$ denotes within-block average variation, and $B = \min_{\{w_{ij} = 0\}} |\hat{\theta}_i - \hat{\theta}_j|$ denotes the between-block separation. A larger $B$ and smaller $W$ indicate a better block identification performance. The results are summarized in Table 1.

Table 1: Comparison among Bayesian t-fusion, Bayesian Laplace fusion, and $L_1$ fusion. The report are based on average results from 100 replications. Refer to Section 4.1 for the definition of $R$, $W$ and $B$.

| Methods                      | Bayesian t-fusion | Bayesian Laplace fusion | $L_1$ fusion |
|------------------------------|-------------------|-------------------------|--------------|
| $L_2$ error of $\theta$      | 1.3243            | 2.2323                  | 1.5916       |
| Standard error               | 0.0602            | 0.0324                  | 0.0421       |
| $L_1$ error of $\theta$      | 9.6964            | 16.6296                 | 11.6745      |
| Standard error               | 0.3419            | 0.2725                  | 0.3407       |
| $W$ (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 simulation results show that the Bayesian t fusion yields the most accurate estimation in terms of both $L_2$ and $L_1$ errors. 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, but unlike the LASSO penalty which introduces a bias of $\lambda$, fuse-$L_1$ penalty introduces only a bias as small as $\lambda/n_i$ 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 $\hat{\theta}_i$'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) 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 (9) and (10), Dirichlet process prior (13) and the frequentist bCARDS estimation (12) using $L_1$ fusion [20]. Besides the comparison of $L_2$ and $L_1$ errors 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 $\theta$, 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 [37], $E(z_1) \approx 1.04 > 0.96 \approx E(z_2)$. Therefore, misclassification will always happen for those extreme observations, and $B = \min\{w^*_{ij}=0\} |\hat{\theta}_i - \hat{\theta}_j|$ 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

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

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 [10] gives a slightly worse point estimator. But in terms of the
Table 2: Comparison among Bayesian t-shrinkage, Bayesian DP prior, and $L_1$ fusion. The report are based on average results from 100 replications. Refer to Section 4.2 for the definition of $R$, $W$ and $\tilde{B}$.

| Methods                  | posterior mean of $L_2^2$ error | Standard error | $L_2$ error of $\theta$ | Standard error | $L_1$ error of $\theta$ | Standard error | $W$ (the smaller the better) | Standard error | $\tilde{B}$ (the larger the better) | Standard error | $R$ (the smaller the better) | Standard error |
|--------------------------|---------------------------------|----------------|--------------------------|----------------|--------------------------|----------------|-------------------------------|----------------|-----------------------------------|----------------|-------------------------------|----------------|
| $t$-shrinkage (10)       | 20.559                          | 0.6885         | 4.1873                   | 0.0872         | 28.1099                  | 0.7385         | 0.3517                         | 0.0093         | 1.4652                            | 0.0218         | 1752.24                        | 31.5641        |
| $t$-shrinkage (9)        | 20.5660                         | 0.4589         | 4.3925                   | 0.0531         | 33.4543                  | 0.4845         | 0.4428                         | 0.0063         | 1.3392                            | 0.0147         | 2294.76                        | 14.6102        |
| DP prior (13)            | 40.7029                         | 3.1209         | 3.9584                   | 0.1621         | 26.8122                  | 1.7122         | 0.2183                         | 0.0061         | 1.3885                            | 0.0419         | 43.8450                        | 0.7909         |
| $L_1$ fusion (12)        | —                               | —              | 5.0039                   | 0.0739         | 43.8450                  | 0.7909         | 0.4551                         | 0.0062         | 0.9855                            | 0.0146         | 1973.7                         | 11.9937        |

posterior contraction, both $t$-shrinkage approaches have much smaller posterior mean of squared $L_2$ error. This result is consistent to the insight obtained from the toy example in Section 3 that LP prior leads to larger posterior variance. 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 (12) has the worst performance in almost every aspect. In summary, DP prior does yield the best Bayes point estimation, but the adaptive $t$-shrinkage (10) 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 (10) and (9), we conclude that the $r$-update step implemented in algorithm (10) does improve the performance of $t$-shrinkage in every aspect for this clustering problem.
4.3 Real data set analysis

In this section, we consider a real comparative genomic hybridization (CGH) dataset [46]. 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 both Bayesian t-shrinkage fusion [5] and frequentist $L_1$ fusion [2]. For the Bayesian application, the hyperparameter is chosen as $a_t = 2$ and $b_t = 0.0007$, following the suggestion of [7]. Both point estimations are plotted in Figure 5. Frankly speaking, the two estimations share similar pattern, and it is difficult to judge which one is definitely better. To see the slight advantage of t-shrinkage, a closer comparison is displayed in the same figure as well, where we compare the two estimators within the segment [80,120]. It is obvious that both approaches induce exactly the same blocking structure. But the difference is that, for t-shrinkage prior, the $\hat{\theta}_i$ within each block is very close to the sample mean of the block; but for $L_1$ fusion, there is a large bias between $\hat{\theta}_i$ and the sample mean of the block.

![Figure 5: Real data application: 1) CGH data; 2) posterior mean of t shrinkage fusion; 3) frequentist $L_1$ fusion estimator; 4) Comparison of the two approaches within a segment of [80,120].](image-url)
5 Final remarks

In this work, we study the Bayesian inference for a vector parameter $\theta$ 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 $\theta$.

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 $\hat{r}$, and then fuse all pairs of neighbors (that are determined by $\hat{r}$) via shrinkage priors. An adaptive $r$-update step is further incorporated to improve the clustering performance. Simulations show that the proposed algorithms (9) and (10) have reasonable performance, and their posterior contraction is much better than conventional Dirichlet process modeling. Further theoretical investigation is 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 cannot rank multivariate vectors, the possible alternative is to construct a minimum spanning tree (MST)[26], and impose shrinkage fusion prior on the pair of $\theta_i$ that is connected by a edge in the MST.

Acknowledgement

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).
A Appendix

First, let us state some useful lemmas.

**Lemma A.1** (Lemma 1 of [25]). Let \( \chi^2_d(\kappa) \) be a chi-square distribution with degree of freedom \( d \), and noncentral parameter \( \kappa \), then we have the following concentration inequality

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

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

**Lemma A.2** (Theorem 1 of [54]). Let \( X \) be a Binomial random variable \( X \sim B(n,v) \). For any \( 1 < k < n - 1 \)

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

where \( \Phi \) 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 [2]:

**Lemma A.3.** Let \( f^* \) be the true probability density of data generation, \( f_\theta \) be the likelihood function with parameter \( \theta \in \Theta \), and \( E^* \), \( E_\theta \) denote the corresponding expectation respectively. Let \( B_n \) and \( C_n \) be two subsets in parameter space \( \Theta \), and \( \phi_n \) 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,

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

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

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

**Proof.** Define \( \Omega_n \) 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

\[ E^* \pi(C_n \cup B_n)|D_n) = E^* \pi(C_n \cup B_n)|D_n)(1 - \phi(D_n))1_{\Omega_n} \]

\[ + E^* \pi(C_n \cup B_n)|D_n)(1 - \phi(D_n))(1 - 1_{\Omega_n}) + E^* \pi(C_n \cup B_n)|D_n)(1 - 1_{\Omega_n}) + E^* \phi(D_n) \]

\[ \leq E^* \pi(C_n \cup B_n)|D_n)(1 - \phi(D_n))1_{\Omega_n} + E^*(1 - 1_{\Omega_n}) + E^* \phi(D_n) \]

\[ \leq E^* \{m(D_n, C_n \cup B_n)/a_n f^*(D_n))(1 - \phi(D_n)) + b_n' + a_n' \].

By Fubini theorem,

\[ E^* (1 - \phi(D_n))m(D_n, C_n \cup B_n)/f^*(D_n) = \int_{C_n \cup B_n} \int_{\mathcal{X}} [1 - \phi(D_n)] f_{\theta}(D_n) dD_n \pi(\theta) d\theta \]

\[ \leq \int_{C_n} \int_{\mathcal{X}} E_{\theta}(1 - \phi(D_n)) \pi(\theta) d\theta + \int_{B_n} \int_{\mathcal{X}} f_{\theta}(D_n) dD_n \pi(\theta) d\theta \leq b_n + c_n. \]

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 \( \theta \) as \( \{\{1, \ldots, g_1\}, \{g_1 + 1, \ldots, g_2\}, \ldots, \{g_d + 1, \ldots, n\}\} \). Given \( G \) and its corresponding blocking structure, \( \hat{\theta}_G(y) \) denotes the estimation of \( \theta \) based on block mean, i.e. \( \hat{\theta}_G(y) = \sum_{i=g_j+1}^{g_{j+1}} y_i/(g_{j+1} - g_j) \) for all \( g_j + 1 \leq j \leq g_{j+1} \), and \( \hat{\sigma}_G^2(y) = ||y - \hat{\theta}_G||^2/(n - |G| - 1) \).

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

\[ \phi(y) = 1\{|\hat{\theta}_G - \theta^*|| \geq \sqrt{n}\sigma^* \epsilon_n \text{ and } |\hat{\sigma}_G^2 - \sigma^{*^2}| > \sigma^{*^2} \epsilon_n \}

\[ \text{for all } G \supseteq G^*, |G| \leq (1 + \delta)|G^*| \] \hspace{1cm} (14)

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

\[ C_n = \{\theta : ||\theta - \theta^*|| \leq M\sqrt{n}\sigma^* \epsilon_n, (1 - \epsilon_n)/(1 + \epsilon_n) < \sigma^2/\sigma^{*^2} < (1 + \epsilon_n)/(1 - \epsilon_n)\} \setminus B_n. \]

\[ B_n = \{\theta : \text{Among all } \{\theta_i\}_{i \in G^*}, \text{ there are at least } \delta|G^*| \text{ of them are greater than } \sigma \epsilon_n/n\}. \]

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

\[ P(||\hat{\theta}_G(y) - \theta^*|| \geq \sqrt{n}\sigma^* \epsilon_n \text{ and } |\hat{\sigma}_G^2 - \sigma^{*^2}| > \sigma^{*^2} \epsilon_n \) \leq \exp\{-c_1 n \epsilon_n^2\}. \]
for some constant \( c_1 \), since \( |G| = O(|G^*|) < n\epsilon_n^2 \) and \( \epsilon_n \leq 1 \). Therefore,
\[
E_{(\theta, \sigma^2)} \phi(y) \leq \left( \frac{n-1}{(1+\delta)|G^*|} \right) \exp\{-c_1n\epsilon_n^2\} = \exp\{-c'_1n\epsilon_n^2\},
\]
(15)
as long as \( n\epsilon^2/|G^*| \log n \) is sufficiently large.

For any \((\theta, \sigma^2) \in C_n\) satisfying \( \|\theta - \theta^*\| \leq M \sqrt{n}\sigma^* \epsilon_n \) and \( \sigma^2/\sigma^*^2 \geq (1 - \epsilon_n)/(1 + \epsilon_n) \), we define
\[
\hat{G} = \{i : \theta_i - \theta_{i-1} \geq \sigma_n \epsilon_n / \sqrt{n} \} \cup G^* \quad \text{(hence \( |G| \leq (1 + \delta)|G^*| \)),}
\]
thus
\[
P_{(\theta, \sigma^2)}(\|\hat{G} \setminus G^*\| - \theta^* \leq \sqrt{n}\sigma^* \epsilon_n) = P_{(\theta, \sigma^2)}(\|\hat{G} \setminus G^*\| - \theta^* \leq \sqrt{n}\sigma^* \epsilon_n)
\]
\[
\leq P_{(\theta', \sigma^*)}(\|\hat{G} \setminus G^*\| - \theta^* \leq \sqrt{n}\sigma^* \epsilon_n)
\]
\[
\leq P_{(\theta', \sigma^*)}(\|\hat{G} \setminus G^*\| - \theta^* \leq \sqrt{n}\sigma^* \epsilon_n)
\]
\[
\leq P_{(\theta', \sigma^*)}\left( \frac{\sqrt{n\epsilon_n}}{\epsilon_n} (M - 1) - \frac{\epsilon_n^2}{\epsilon_n} \right) \leq \exp\{-c_2n\epsilon_n^2\}
\]
for some \( c_2 \) given a large \( M \), where the second inequality is due to the fact that \( \|\hat{G} \setminus G^*\| - \theta^* \leq \sqrt{n}\sigma^* \epsilon_n \)
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 \( \sigma^2/\sigma^*^2 > (1 + \epsilon_n)/(1 - \epsilon_n) \),
\[
P_{(\theta, \sigma^2)}(\|\hat{G} \setminus G^*\| < \sigma^2 \epsilon_n) = P_{(\theta, \sigma^2)}(\|\hat{G} \setminus G^*\| < \sigma^2 \epsilon_n)
\]
\[
\leq P_{(\theta, \sigma^2)}(1 - \epsilon_n < \|\hat{G} \setminus G^*\| < 1 + \epsilon_n)
\]
\[
\leq P_{(\theta, \sigma^2)}(\|\hat{G} \setminus G^*\| < \sigma^2 \epsilon_n)
\]
\[
\leq P_{(\theta, \sigma^2)}(\|\hat{G} \setminus G^*\| < \sigma^2 \epsilon_n)
\]
\[
\leq P_{(\theta, \sigma^2)}\left( |\lambda^2_{n-|G^*|-1}(\lambda) - (n - |G^*| - 1)\epsilon_n | \right) \leq \exp\{-c'_2n\epsilon_n^2\}
\]
for some \( c'_2 \), where the noncentral parameter \( \lambda < n\epsilon_n^2 < (n - |G^*| - 1)\epsilon_n \).

Combining the results from the previous two paragraphs, it is easy to obtain that
\[
\sup_{(\theta, \sigma^2) \in C_n} E_{(\theta, \sigma^2)}[1 - \phi(y)] \leq \max\{\exp(-c_2n\epsilon_n^2), \exp(-c'_2n\epsilon_n^2)\}
\]
(16)

Now we consider the marginal posterior density of data \( y \). With probability \( P(\|\epsilon\| \leq 2\sqrt{n}\sigma^*) \) (which converges to 1),
\[
\frac{m(y)}{f^{*}(y)} = \frac{\int_{\theta} \sigma^n \exp \left[ -\|\theta^* - \theta + \epsilon\|^2/\sigma^2 \right] d\theta d\sigma^2}{\sigma^n \exp \left[ -\|\epsilon\|^2/\sigma^*^2 \right]}
\]
\[
\geq \int_{\sigma^2} \int_{\theta} \exp \left\{ -\|\theta^* - \theta\|^2/\sigma^2 - 2 (\theta^* - \theta) \epsilon \frac{\epsilon}{\sigma^2} + \frac{\|\epsilon\|^2}{\sigma^2} - \frac{\|\epsilon\|^2}{\sigma^*^2} - n \log(\sigma/\sigma^*) \right\} \pi(\theta, \sigma^2) d\theta d\sigma^2
\]
\[
\geq \pi(\max\{|\theta_i - \theta_{i-1}^*|, |\theta_i - \theta_i^*|\} \leq |G^*| \log n/n^2, 0 \leq \sigma^2 - \sigma^*^2 \leq \sigma^2 |G^*| \log n/n) \exp\{-c_3|G^*| \log n\}
\]
26
for some constant $c'$. Besides,

$$\pi(\max \{ |\theta_i - \hat{\theta}_i|, |d_i - \hat{\theta}_i| \} / \sigma \leq |G^*| \log n / n, 0 \leq \sigma^2 - \sigma^* \leq \sigma^2 G^* \log n / n)$$

$$\geq \pi_o(\sigma^2) \ast O(\sigma^2 G^* \log n / n) * \pi_{0} * O(|G^*| \log n / n^2)$$

$$* \pi_g^{G^*} ||G^*| \log n / n^2 ||G^*| * [\pi_o\{\{-|G^*| \log n / n^2, |G^*| \log n / n^2\}\}]^n$$

$$= \exp\{-c_3' |G^*| \log n\}.$$ (by the conditions imposed on the prior specifications)

for some constant $c'_3$. Thus

$$m(\phi)/f^*(\phi) \geq \exp\{-(c_3 + c'_3)|G^*| \log n\} = \exp\{c_3 n |\epsilon_n|^2\}, \text{ with probability tending to 1}$$ (17)

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 $\theta_i$s, $\pi(B_n) = \pi[\text{Bin}(n - 1 - |G^*|, p) > (\delta)|G^*|]$, where $p \leq (1/n)^{1+u}$. By Lemma A.2

$$\pi(B_n) \leq \exp\{-c_4 \delta |G^*| \log n\}$$ (18)

for some $c_4$. Combine results (15), (16), (17) and (18), and we apply Lemma A.3 to get the posterior consistency result that

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

given a sufficient large constants $\delta$ 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 $f(d) = n! \prod \phi(d_i)1(d_1 \leq d_2, \ldots, d_{n-1} \leq d_n)$ and $\phi$ denotes the standard normal density. The prior of $\theta$ follows $\pi(\theta) = \pi_1(\theta_1) \prod_{i=2}^{n} \pi_{t,s}(\theta_i - \theta_{i-1})$, where $\pi_{\lambda}$ is the density of $N(0, \lambda)$, 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 $1 \leq i \leq k = 3n/4$, 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 $\|\Delta \theta\|^2 \leq M \log n$. 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
\[
\log \left( \frac{\exp\left[-(y - \mu - \Delta \theta)^2/2\right]}{\exp\left[-(y - \theta^* - \Delta \theta)^2/2\right]} \right) = \frac{1}{2} \sum_{i=k+1}^n \left[ (y_i - \Delta \theta_i)^2 - (y_i - Z_{0.25}/2 - \Delta \theta_i)^2 \right] = \frac{1}{2} \sum_{i=k+1}^n \left[ (y_i - \Delta \theta_i) \left( Z_{0.25} - \frac{Z_{0.25}^2}{4} \right) \right]
\]
\[
\geq \frac{1}{2} \left[ \left( \|y_{k+1:n}\|_1 - \sqrt{nM \log n/4} \right) Z_{0.25} - \frac{nZ_{0.25}^2}{16} \right] \geq cn,
\]
for some positive constant \(c\) given sufficiently large \(n\), where the inequalities above hold since \(y_n \geq y_{n-1} \cdots \geq y_{k+1}\, \text{and} \, y_{k+1} \approx Z_{0.25}\) 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 \(\mu\) than at the true parameter \(\theta^*\).

**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:

\[
C_n = \{ \theta : \|\theta - \theta^*\| \leq M \sqrt{n} \sigma^* \epsilon_n, (1 - \epsilon_n)/(1 + \epsilon_n) < \sigma^2/\sigma^{*2} < (1 + \epsilon_n)/(1 - \epsilon_n) \} \setminus B_n,
\]

\[
B_n = \{ \theta : \text{Among all } \{\theta_i - \theta_i-1\}_{i=2}^n, \text{there are at least } \delta \text{ of them are greater than } \sigma \epsilon_n/n \}.
\]

Using the same arguments, one can still establish exponential separation results (15) and (16).

To establish (17), we notice that

\[
\frac{m(y)}{f^\star(y)} = \int_{\Theta^2} \int_\Theta \pi^{**} \exp\left\{ -\|\theta^* - \theta + \varepsilon\|^2/\sigma^2 \right\} d\theta d\sigma^2/\sigma^n \exp\left\{ -\|\varepsilon\|^2/\sigma^{*2} \right\}
\]

\[
\geq \int_{\Theta^2} \int_\Theta \exp\left\{ -\|\theta^* - \theta\|^2/\sigma^2 - 2 \left( \theta^* - \theta \right)^T \varepsilon \sigma/\sigma^2 + \|\varepsilon\|^2/\sigma^{*2} - n \log(\sigma/\sigma^*) \right\} \pi(\theta, \sigma^2) d\theta d\sigma^2
\]

\[
\geq \pi(\max[\|\theta_i - \theta^*_i\|/\sigma \leq \log n/n, 0 \leq \sigma^2 - \sigma^{*2} \leq \sigma^{*2} \log n/n]) \exp\{-c_3 \log n\}
\]

for some constant \(c_3\) and

\[
\pi(\max[\|\theta_i - \theta^*_i\|/\sigma \leq \log n/n, 0 \leq \sigma^2 - \sigma^{*2} \leq \sigma^{*2} \log n/n, 0 \leq \sigma^2 - \sigma^{*2} \leq \sigma^{*2} \log n/n]) \pi(r).
\]

This ensures (17).

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 \(\int_{k \epsilon_0/n} \pi_{t,s}(x)dx \geq 1 - 1/n^{1+w}\) 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 $-\log(\pi(B_n)) \geq u \log n$.

References

[1] David F Andrews and Colin L Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102, 1974.

[2] A. Barron. Information-theoretic characterization of bayes performance and the choice of priors in parametric and nonparametric problems. In J.M. Bernardo, J. Berger, A. Dawid, and A. Smith, editors, *Bayesian Statistics 6*, pages 27–52, 1998.

[3] James O Berger, Xiaojing Wang, and Lei Shen. A bayesian approach to subgroup identification. *Journal of biopharmaceutical statistics*, 24(1):110–129, 2014.

[4] Brenda Betancourt, Abel Rodríguez, and Naomi Boyd. Bayesian fused lasso regression for dynamic binary networks. *Journal of Computational and Graphical Statistics*, 26(4):840–850, 2017.

[5] Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet-laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110:1479–1490, 2015.

[6] C.M. Carvalho, N.G. Polson, and J.G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480, 2010.

[7] I. Castillo, J. Schmidt-Hieber, and A.W. van der Vaart. Bayesian linear regression with sparse priors. *Annals of Statistics*, pages 1986–2018, 2015.

[8] Ismaël Castillo and Aad van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101, 2012.

[9] J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771, 2008.
[10] J. Chen and Z. Chen. Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 22:555–574, 2012.

[11] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[12] Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

[13] Subhashis Ghosal and Aad W Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, 35(1):192–223, 2007.

[14] P.R Hahn and C.M. Carvalho. Decoupling shrinkage and selection in bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, 110:435–448, 2015.

[15] Katherine A Heller and Zoubin Ghahramani. Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, pages 297–304, 2005.

[16] H. Ishwaran and J.S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, pages 730–773, 2005.

[17] W. Jiang. Bayesian variable selection for high dimensional generalized linear models: Convergence rate of the fitted densities. *Annals of Statistics*, 35:1487–1511, 2007.

[18] V.E. Johnson and D. Rossel. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107:649–660, 2012.

[19] Iain M Johnstone. High dimensional bernstein-von mises: simple examples. *Institute of Mathematical Statistics collections*, 6:87, 2010.

[20] Zheng Tracy Ke, Jianqing Fan, and Yichao Wu. Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509):175–194, 2015.

[21] Zheng Tracy Ke, Jianqing Fan, and Yichao Wu. Homogeneity pursuit. *Journal of the American Statistical Association*, 110:175–194, 2015.
[22] Bas JK Kleijn, Aad W van der Vaart, et al. Misspecification in infinite-dimensional bayesian statistics. *The Annals of Statistics*, 34(2):837–877, 2006.

[23] B.J.K. Kleijn and A.W. van der Vaart. Misspecification in infinite-dimensional bayesian statistics. *Annals of Statistics*, 34:837–877, 2006.

[24] Minjung Kyung, Jeff Gill, Malay Ghosh, and George Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

[25] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[26] Furong Li and Huiyan Sang. Spatial homogeneity pursuit of regression coefficients for large datasets. *Journal of the American Statistical Association*, 2018.

[27] H. Li and D. Pati. Variable selection using shrinkage priors. *Computational Statistics & Data Analysis*, 107:107119, 2017.

[28] F. Liang, Q. Song, and K. Yu. Bayesian subset modeling for high dimensional generalized linear models. *Journal of the American Statistical Association*, 108:589–606, 2013.

[29] Jun Liu, Lei Yuan, and Jieping Ye. 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*, pages 323–332, 2010.

[30] Shujie Ma and Jian Huang. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423, 2017.

[31] Alexander Mozeika and Anthony Coolen. Mean-field theory of bayesian clustering. *arXiv preprint arXiv:1709.01632*, 2018.

[32] N.N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

[33] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.
[34] T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008.

[35] Alessandro Rinaldo et al. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B):2922–2952, 2009.

[36] Herbert Robbins. An empirical bayes approach to statistics. In *Herbert Robbins Selected Papers*, pages 41–47. Springer, 1985.

[37] JP Royston. Algorithm as 177: Expected normal order statistics (exact and approximate). *Journal of the royal statistical society. Series C (Applied statistics)*, 31(2):161–165, 1982.

[38] J.G. Scott and J.O. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, pages 2587–2619, 2010.

[39] Xiaotong Shen and Hsin-Cheng Huang. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105:727–739, 2012.

[40] Kaito Shimamura, Masao Ueki, Shuichi Kawano, and Sadanori Konishi. Bayesian generalized fused lasso modeling via neg distribution. *Communications in Statistics-Theory and Methods*, pages 1–23, 2018.

[41] Q. Song and F. Liang. A split-and-merge bayesian variable selection approach for ultra-high dimensional regression. *Journal of the Royal Statistical Society, Series B*, in press, 2014.

[42] Qifan Song and Faming Liang. Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*, 2017.

[43] X. Tang, X. Xu, M. Ghosh, and P. Ghosh. Bayesian variable selection and estimation based on global-local shrinkage priors. *arXiv:1605.07981*, 2016.

[44] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.
Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and
smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Method-
ology)*, 67(1):91–108, 2005.

Robert Tibshirani and Pei Wang. Spatial smoothing and hot spot detection for cgh data using the
fused lasso. *Biostatistics*, 9(1):18–29, 2007.

Sara van der Geer and Peter Bühlmann. *Statistics for High-Dimensional Data: Methods, Theory
and Applications*. Spring Series in Statistics, Springer, 2011.

S.L. van der Pas, B. Szabo, and Aad van der Vaart. Adaptive posterior contraction rates for the
horseshoe. *arXiv:1702.03698*, 2017.

Sara Wade and Zoubin Ghahramani. Bayesian cluster analysis: Point estimation and credible balls.
*Bayesian Analysis*, 13:559–626, 2018.

Z. Xu, D.F. Schmidt, E. Makalic, G. Qian, and J.L. Hopper. Bayesian sparse global-local shrinkage
regression for grouped variables. *arXiv:1709.04333*, 2017.

Yun Yang, Martin J Wainwright, and Michael I Jordan. On the computational complexity of high-
dimensional bayesian variable selection. *Annals of Statistics*, in press, 2015.

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

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

AM Zubkov and AA Serov. A complete proof of universal inequalities for the distribution function
of the binomial law. *Theory of Probability & Its Applications*, 57:539–544, 2013.